The University of Sheffield

J.E.Brazier

# Valuing health benefits: the development of a preference-based measure of health for use in the economic evaluation of health care from the SF-36 Health Survey

This thesis has been submitted for the degree of Doctor of Philosophy

Vol 1

Year of submission: 1997

## Declaration

The research reported in this thesis was based on a project funded by the Department of Health to derive a single index from the SF-36. The project proposal was designed and written by me. The derivation of the SF-6D health classification is acknowledged to have been a team effort. Two thirds of the valuation data were collected by the research assistant on the project. I undertook all aspects of the data cleaning, manipulation, and the univariate and multivariate analyses.

I wrote this thesis. Any written passages taken from someone elses work has been distinguished by quotation marks, and all sources of information acknowledged.

This work has not been submitted in any previous application for a degree.

John Brazier
May 1997

Acknowledgements

I am very grateful for the advice and support of a great many people in the production of this thesis.

My thanks go to the original team involved in the Department of Health funded project to derive a single index from the SF-36: my fellow applicant Tim Usherwood, who suggested the idea, my colleagues Kate Thomas and Nicola Jones, and most of all the research worker on the project Rosemary Harper, who has since provided considerable help and encouragement in writing of this thesis.

There are many people who have assisted on various aspects of the thesis to whom I am indebted: Andrew Booth, Mark Deverill, Simon Dixon, Paul Dolan, Colin Green, Crispin Jenkinson, Alistair Leyland (who managed to explain multi-level modelling to me), Adam Lowey, Paul Lambert, James Munro, Jon Nicholl, Jenny Roberts, and Phil Shackley. I am especially grateful to Chris McCabe who kindly agreed to read and comment on the penultimate draft and to Ron Akehurst for his support and encouragement to complete the thesis.

As a staff candidate, I was not entitled to supervisors, but have had two advisors who I would like to thank: John Cairns for advice in the early stages of the writing of the thesis, and Peter Else for reading and commenting on the manuscripts.

On the production side I am very grateful to J R Atkins for having the patience to proof read the thesis, to Jane Viney for her assistance and to my partner Cathy for organising and assisting in this endeavour. All partners of Phd students, especially of part time students, deserve recognition for their support and tolerance of the very unreasonable burdens they suffer, and Cathy is no exception. The hidden costs are many! Thanks are also extended to Natalka Kurlak for typing the first draft of the thesis and helping with its production at the end.

Finally I wish to acknowledge the financial support of the Department of Health and the NHS National Executive (Trent). The views expressed in this thesis are those of the author, and not either of these organisations.

# Contents

## Summary

### Valuing health benefits: the development of a preference-based measure of health for use in the economic evaluation of health care from the SF-36 Health Survey

by J.E.Brazier

The main aim of the research was to develop a preference-based measure of health from the Short Form-36 (SF-36) Health Survey for valuing health-related quality of life on a 0 to 1 scale in order to calculate Quality adjusted life years (QALYs).

Before undertaking the empirical work, reviews were undertaken of the justification for the QALY approach, existing preference-based measures for deriving QALYs and the rationale for looking at the SF-36.

The methods of the research were as follows. The SF-36 was reduced and simplified to form a six dimensional health state classification (SF-6D) amenable to valuation. One hundred and sixty five patients, health professionals, managers, and students valued a sample of health states defined by the SF-6D using the visual analogue scale (VAS) and standard gamble (SG) techniques to elicit preferences. There were 1,357 VAS and 1,037 SG health state valuations after adjustment and exclusions for major inconsistencies. Models for predicting median and mean VAS and SG health state values from the SF-6D were estimated from these data by multivariate techniques.

A set of additive models were selected on the basis of goodness of fit and parsimony. More complex specifications did not improve the models. Initial applications of algorithms based on these models to five data sets suggested this new preference-based measure retained much of sensitivity of the SF-36 at the milder end of the of the illness spectrum.

The preference-based algorithms can be used to transform SF-36 data collected in a clinical trial (with costs) into information suitable for assessing the cost-effectiveness of health care interventions. The adoption of these algorithms has the potential to considerably extend the application of economic evaluation in health care.

## Overview

The main aim of the research reported in this thesis was to develop a preference based measure of health from the Short Form-36 (SF-36) Health Survey for valuing health related quality of life on a 0 to 1 scale in order to calculate Quality Adjusted Life Years (QALYs).

Health Economists have long recognised that the main purpose of consuming health care is to promote good health. Unique to health economics has been the development of this notion of good health into the measure of a 'year in full health', which combines length of life with health related quality of life. The most commonly used version of this measure is the Quality Adjusted Life Year (QALY). The number of QALYs is calculated by multiplying a person's life expectancy by the weight assigned to the health related quality of life experienced in each period, ranging from 0 to 1.0, where 0 is assigned to death and 1.0 to full health. This measure is used in cost-utility analysis, where the cost-effectiveness of health care interventions is compared in terms of their cost per QALY.

The thesis begins by examining the theoretical justification for using the QALY measure and the reasons why it has been favoured by many health economists over more conventional monetary measures. The QALY implies the following restrictions to the individuals utility function: it excludes non-health benefits; the time preference rate for health is zero; the value of a health state is independent of the time spent in the state; the value of a health state is independent of the health state(s) which went before it or are expected to come after it and the individual has a constant attitude to risk. These assumptions make the QALY a very versatile measure, and one which can be used in decision tree analysis and markov modelling. They have been criticised in the literature for misrepresenting preferences for health care, but there is little evidence regarding the significance of the violations for decision making. The alternative measures of the Healthy Year Equivalent and *ex ante* QALYs are more difficult and inflexible to apply and a more complex and lengthy set of valuation tasks. These alternatives have involved

major simplifications in their descriptions of the health scenarios of health care interventions.

There are currently five preference-based measures of health used to estimate QALYs, namely: the Quality of Well-Being scale (QWB), Rosser's disability/distress scale, the Health Utility Index versions one, two and three (HUI-I, HUI-II, and HUI-III), the EQ-5D (EuroQoL[c]), and the 15D. These were systematically reviewed against the criteria of practicality, reliability, descriptive validity (content and construct validity), validity of the preference weights, and empirical validity (against revealed, stated, and hypothetical preferences). This review was undertaken using papers identified by a systematic search of the literature. The literature search found 163 papers on these instruments. Seventy one of these papers were applications of the measures that provided the empirical evidence.

The review found the EQ-5D to be the best preference based measure of health in adult populations. This conclusion would have to be re-appraised when (Canadian) weights become available for the HUI-III. In the UK, researchers are likely to continue to favour the EQ-5D on the grounds that it has been more widely used in this country and there are a set of algorithms obtained from a large representative survey of adults in the UK using the time trade-off technique of preference elicitation. The dimensions of the EQ-5D cover most dimensions of general health, but the three levels per dimension of health would on the face of it seem too crude to detect smaller changes. There is little evidence on descriptive validity. What is available suggests it can detect large differences, though there is evidence of its insensitivity.

There is a case for examining the potential for developing a larger and more sensitive preference based measure of health than the EQ-5D. The question is whether the finer differences described by a larger classification would be important in terms of preferences.

The SF-36 health survey is a brief self-completed questionnaire which generates scores across eight dimensions of health. The SF-36 is an important measure of general health

and one of the most commonly used in clinical trials in the UK, the rest of Europe and North America. The review of its use presented in chapter 4 found it to be practical in terms of its ease of use, achieving high levels of response and completion, and to be reliable. The strength of this measure lies in its descriptive validity, and in particular its sensitivity. Evidence was found of its greater sensitivity compared to the Rosser and the EQ-5D at detecting milder conditions and at responding to health changes in some groups of patients.

The SF-36 is potentially a rich source of data for economic evaluation, but has only a limited use in assessing cost-effectiveness because the scores are not based on preferences. Dimension scores are computed by adding item responses together assuming equal weighting. It is impossible to evaluate the relative cost-effectiveness of interventions when trade-offs must be made between dimensions of the SF-36, between these dimensions and survival, and cost. SF-36 dimension scores could be incorporated into the framework of a cost-consequences analysis, but this would be of limited help to decision-makers given the difficulties of interpreting the scores. The incorporation of preference values into SF-36 in order to be able to derive health state values for calculating QALYs would considerably extend the application of cost-utility analysis in health care.

Four approaches to incorporating preferences into the SF-36 were considered: to map items of the SF-36 onto an existing preference based measure; to estimate exchange rates for converting these scores into preference values; to construct vignettes from the results of each trial and to value these using one of the preference elicitation techniques; or to value a multi-dimensional scale based on the SF-36. The last approach has been chosen since it would produce a measure based on preferences that can be used in more than one economic evaluations, and the results can be used to inform resource allocation decisions between programmes as well as within a patient group.

The chosen approach has three components. The first component is to reduce and simplify the content of the SF-36 to form a multi-dimensional health state classification suitable for valuation. The second part is to value a sample of health states defined by

the classification. The third part is to estimate values or weights for multi-dimensional classification from the sample of health states in order to be able to value all possible health states defined by the classification.

The adaptation of the SF-36 for valuation was undertaken by a multi-disciplinary research team based in Sheffield led by the author. At a series of meetings the team collectively arrived at decisions based on the following judgement criteria: to avoid duplication of items; to exclude positive items; and to use the views of patients, health professionals and members of the general public where available. The result was a six dimensional classification called the Short-Form Six-Dimensional Health State classification (SF-6D). The dimensions had between two and six levels. This classification defines are a total of 9000 health states. All responders to the SF-36 questionnaire can be assigned to the SF-6D provided the 14 items used in the classification have been completed.

A sample of 57 health states (excluding full health) were chosen for the valuation survey. They were selected to provide a balance of states in terms of physical and mental problems and severity. To ensure all health states were plausible, the selection was limited to those which occurred in existing SF-36 data sets. These health states were valued by a convenience sample of 165 patients, health professionals, health managers, and students, who were asked to undertake three valuation tasks: ranking, rating on a visual analogue scale, and valuation by standard gamble (SG). The patient sub-sample valued eight health states in this way and the non-patients valued 12.

SG had been chosen as the main technique for eliciting preferences, since risk attitude towards health status is incorporated through the elicitation of utility values under conditions of uncertainty. Furthermore, an attraction of SG is that it mirrors elements of medical decisions. On these grounds, SG was preferred to other choice based techniques, such as Time trade-off (TTO).

A self-completed version of SG was selected on grounds of practicality. This version has been found to be no worse than Visual analogue scale (VAS) or TTO in terms of

consistency and reliability. The SG question, however, had to be adapted for the valuation of the milder states defined by the SF-6D. The VAS was primarily included to familiarise the respondent with health state valuation.

The respondents to the survey were not a representative sample of any one group, but nonetheless reflected a range of backgrounds and illness experiences. The quality of the VAS and SG data in terms of the rates of completion and consistency compared favourably with other surveys, though there was evidence of instability in the valuations from the split test.

The 165 respondents provided 1,582 VAS ratings and 1,567 SG values for the 57 health states. VAS data were adjusted by transforming the results onto a scale of 1.0 for health state 111111 (i.e full health) and zero for death. After this adjustment, and the exclusions for major inconsistencies with the SF-6D classification, there were 1,357 VAS observations by 155 respondents. The main exclusions from the SG data were those gambles with a non-fatal outcome reference state because these were found to produce values that were inconsistent with those obtained from gambles with death as the worse reference state. All patient respondents, who mainly undertook non-fatal gambles, have therefore been excluded. This left 1,037 SG observations from 106 respondents.

There are practical benefits from being able to use VAS instead of SG to value health states because it is easier to complete, more reliable, and results in lower respondent confusion. There has also been considerable theoretical interest in the relationship between these two. It was therefore decided to attempt to estimate the relationship between VAS and SG at the individual and health state level.

SG values were found to exceed VAS ratings and the plot between the two had a characteristic bowing outwards in a northward direction. The conventional explanation for this relationship was that the difference between VAS and SG is a persons (constant) attitude to risk and a person who was risk averse would exhibit this concave relationship. This would imply a power function. However, there was also evidence of a

positive non-zero intercept, lending support to the alternative Gambling effect hypothesis. Furthermore, there are competing explanations for the concave power function. At the individual level, a range of model specifications were found to fit the data poorly, and there was evidence of non-normal residuals and heteroscedasticity. A better relationship was found between VAS and SG for mean health state values, but the parameters of the models were different to those found in other studies. There is therefore no theoretical or empirical support for transforming VAS scores into SG utilities at the aggregate or individual level. The implications for the research presented in this thesis is that the modelling of health state values must be undertaken with actual SG data rather than values extrapolated from the VAS data.

The aim of the multivariate analysis was to find the best models for predicting VAS and SG values for health states defined by the SF-6D. Models were estimated for both mean and median values of these variables owing to their skewed distribution. An additive specification was used, with the dimension levels of the SF-6D entered as dummy independent variables. For the median models, the unit of analysis was the 57 average health state values and weighted least squares used as the technique of estimation. The unit of analysis for the mean models was the individual valuations, since this made better use of the data. A problem with a pooled panel data set is that observations are not independent, thus violating one of the assumptions of ordinary least Square. Therefore, a fixed-effects adjustment was made for between respondent variation and this was found to substantially improve the fit of the model.

The additive specification was able to explain much of the variation, with adjusted R-squareds of 0.96 for the VAS median, 0.68 for the VAS (individual) mean, 0.97 for the SG median, and 0.49 the SG mean. The first three models passed the standard diagnostic tests (normality of residuals, heteroscedasticity, and overall specification), but the SG mean had problems in terms of non-normality in its residuals and heterogeneity, though it passed in the general test of mis-specification. These problems was not resolved by running the model on a logit transformation of the SG values. Similar problems were encountered in the modelling of health state values for the EQ-

5D from TTO data. It was reassuring to demonstrate the robustness of the parameter estimates of all four models by a split sample test.

There was inconsistency between some coefficients of adjacent dimension levels with the logical ordering of the scales of the SF-6D. This was probably due to multi-collinearity between dimensions. It was necessary to merge some adjacent levels to remove these inconsistencies, and this had the effect of reducing the size of the model. The addition of interaction terms was not found to improve the fit of the model. More complex specifications of between respondent variation at the individual level using a multi-level modelling package improved the goodness of fit, but did not substantively change the coefficients on the levels. Consistent additive versions of the models have been selected to provide the algorithms for valuing the SF-6D.

The algorithms were applied to five patients SF-36 data sets: general population, elderly women, chronic obstructive pulmonary disease, osteoarthritis of the knee, and hernia repair. The primary purpose was to examine the extent to which the adaptation of the SF-36 into the SF-6D and the further simplifications brought about by the modelling (i.e. the merging of dimension levels) reduced the sensitivity of the original instrument to health differences and changes. The results were examined in terms of the reliability, descriptive validity, and empirical validity of the values generated from the SF-6D, and a comparison was undertaken with the EQ-5D. The algorithms were also used to undertake a cost-utility analysis using the results of a randomised clinical trial of alternative treatments for inguinal hernia patients.

There was some evidence of a loss of sensitivity compared to the original SF-36 dimension scores, particularly in terms of responsiveness to health change. The loss was partly a result of the scoring algorithm for deriving the single value, which pools the changes across dimensions. The apparent reduction in responsiveness may reflect the strength of peoples preferences for the overall change and not simply those changes that occur for one or two of the dimensions.

Despite the reduction in sensitivity, the SF-6D values has retained some of the advantages of the SF-36 over the EQ-5D in terms of descriptive validity at the milder end of the spectrum of illness. It was found, for example, that the SF-6D values were able to detect perceived health changes in chronic obstructive pulmonary disease (COPD) patients that was missed by the EQ-5D. There were too few studies, however, to be conclusive about the extent and generalisability of any advantage. The evidence on empirical validity against stated preferences was also inconclusive, since it was limited to own VAS ratings, which are subject to contextual effects.

Mean health state values have been calculated for the five patient data sets by the SF-6D and EQ-5D. The values generated by these measures were found to rank the five groups in the same order i.e. the general population has the highest values, followed by hernia repair, elderly female, COPD, and osteoarthritis of the knee. However, the size of the health state values and the intervals between the mean health state values of the samples were very different. This has important implications for predicting patient choice, evaluating the cost-effectiveness between alternatives for the same patient groups by cost-utility analysis, and for making cross-programme comparisons in terms of cost per QALY.

The research has been successful in estimating a set of preference based algorithms for valuing the SF-36. The application of the SG algorithm to the trial of treatments for inguinal hernia demonstrated how the otherwise ambiguous SF-36 and cost results were transformed into information suitable for assessing the cost-effectiveness of health care interventions. The primary purpose of this research has been achieved. The adoption of the algorithms has the potential to considerably extend the application of economic evaluation in health care. Furthermore, it provides an alternative to existing preference based measures for estimating QALYs and it may prove to be more suitable in some circumstances, particularly for milder conditions. There is considerable scope, however, for further research to improve this new preference based measure by revising the classification, undertaking larger valuation surveys and improving the econometric modelling.

# Chapter 1

# Introduction

## 1.1. Background

*"Good health is one of man's most precious assets. The desire to live, to be well, to maintain full command over one's faculties and to see one's loved ones free from disease, disability or premature death are amongst the most strongly rooted of all human desires."* Fuchs (1966)

Health economists have long recognised that the main purpose of consuming health care is to promote good health (Feldstein, 1963; Fuchs, 1966; Culyer, 1971a; Grossman, 1972). Unique to health economics has been the development of this notion of good health into the measure of a 'year in full health', which combines length of life with health-related quality of life. The most commonly used version of this measure is the Quality Adjusted Life Year (QALY), defined as *'a measure of health outcome which assigns to each period of time a weight, ranging from 0 to 1, corresponding to the health-related quality of life during that period, where a weight of 1 corresponds to optimal health, and a weight of 0 corresponds to a health state judged to be equivalent to death'* ( P.405 Gold et al., 1996). The number of QALYs is calculated by multiplying a person's life expectancy by the weight assigned to the health-related quality of life experienced in each period. The health-related quality of life associated with hospital renal dialysis, for example, may be assigned a weight or quality adjustment value of 0.8. A 20 year period on renal dialysis is therefore 16 QALYs, which is assumed to be equivalent to someone living for 16 years in full health. For more complex health profiles, involving transitions between states of health, the QALY score is calculated by summing the product of the time spent in each state and their quality adjustment value (on a zero to 1.0 scale).

There are two components to the process of estimating the quality adjustment value. The first is to describe a person's state of health and the second is to value that

description. A common approach to estimating the quality value is to administer a standardised questionnaires in a clinical trial to patients to describe their general state of health at various points in time. These questionnaires come with a set of 'off-the-shelf' preference weights from a valuation survey, usually of the general population, using one of a number of preference elicitation techniques. There are currently five of these preference-based measures of health, each with a different selection of dimensions and items for describing health. These measures are intended to be general and relevant to all medical conditions.

The QALY has been used to compare the cost-effectiveness of health care interventions in terms of their marginal cost per QALY gained within and between disease groups (Williams, 1985; Torrance and Zipursky, 1984; Kaplan and Bush, 1982). Such information is potentially useful in assisting purchasers of health care to obtain the maximum health gain from any given budget.

The preference-based measures used to derive the quality adjustment value have been criticised for their descriptions being too insensitive or irrelevant to the health experiences of the patients (Donaldson et al., 1988; Laupacis, 1990; Carr-Hill and Morris, 1991; Revicki and Kaplan, 1993; Brazier et al., 1993; Katz et al., 1994). There is a reluctance on the part of many clinical researchers to use these QALY instruments in clinical trials (Drummond and Davies, 1991). These measures have only been used in a very limited way in clinical trials of new health technologies (Backhouse et al., 1992), and are not sufficiently employed to provide an up-to-date assessment of the cost-effectiveness of interventions.

By contrast, non preference-based measures of health status such as the Short Form-36 (SF-36) Health Survey or the Sickness Impact Profile, tend to be more highly regarded amongst clinical researchers in terms of their relevance and sensitivity (Bowling, 1991; Wilkin et al., 1992; McDowell and Newell,1987). These measures have become used extensively in clinical trials and are an important source of qualitative data regarding the benefits of health care. Some health economists have attempted to use them in economic evaluations alongside clinical trials in order to conduct cost-consequences analysis (e.g.

Buxton et al., 1985; Nicholl et al.,1992). However, the use of such health measures in economic evaluation has been criticised by economists, largely because they do not explicitly incorporate preferences (Culyer, 1978; Williams, 1992; Johannesson et al., 1996) and because they are of limited usefulness in economic evaluation (Brazier,1995).

## 1.2 Aims

The main purpose of the research reported in this thesis is to develop a way of incorporating preferences into the Short Form-36 (SF-36) Health Survey, one of the most commonly used measures of health status, in order that it can be employed to estimate the quality adjuster for calculating QALYs. Within this overall aim, there are a number of specific objectives to be addressed in the thesis:

- to identify the key methodological issues in adapting health measures for use in economic appraisal,
- to change the SF-36 to make it amenable to valuation,
- to undertake a survey to elicit preferences for health states defined by the SF-36,
- to select and apply the appropriate econometric techniques for estimating an algorithm for valuing the SF-36, and
- to determine the extent to which the changes to the SF-36 has reduced the sensitivity of the original instrument to health differences and changes.

## 1.3 Structure and content of thesis

The research reported in the thesis is based on the application of the QALY approach to valuing health care benefits for economic evaluation and hence it is important to establish the economic foundation of this approach. Chapter 2 therefore begins by examining the reasons for not using the more conventional monetary measures of the

benefits (of health care) more commonly used in other areas of economics. The chapter then outlines the key features of the QALY, including its focus on health benefits, the assumptions it makes about people's preferences and its use in informing public decisions. The chapter also addresses the criticisms of the QALY approach, the recently proposed alternative of Healthy Year Equivalents and the reasons for deciding to use the QALY approach to valuing the SF-36.

An important justification for this research is the limitation of existing methods for estimating the quality adjustment value. Chapter 3 reviews existing methods. It begins by presenting the different methods for estimating the quality value, and the advantages of preference-based measures using standardised questionnaires (for obtaining the descriptive data about health-related quality of life) over the use of condition-specific scenarios, or direct utility assessment. This is followed by a review of techniques for eliciting people's preferences. This section provides an important input into the design of the valuation survey presented later in the thesis. The next section presents a systematic review of the five preference-based health measures used to derive the quality adjuster: the Quality of Well-Being scale, Rosser's Disability and Distress scale, the Health Utility Index versions one two and three (HUI-I, HUI-II and HUI-III), the EQ-5D (formerly the Euroqol) and the 15D. This section includes a detailed description of the instruments and a review against the criteria of practicality, reliability, descriptive validity, appropriateness of the valuation of the instrument and the properties of the scores derived from the instrument in terms of empirical validity. This review of existing instruments reveals their strengths and weaknesses, and the rationale for considering other measures of health, such as the SF-36.

Chapter 4 presents a critical overview of the subject of the thesis, the SF-36 health survey. A description of this health measure is followed by a section outlining the case for using it in economic evaluation by reviewing the evidence for its reliability and descriptive validity, including its apparent sensitivity to more mild health problems. The chapter then sets out the economic criticisms of the SF-36 and the limited circumstances in which it can be used to assess the cost-effectiveness of alternative interventions. The

final section examines how the SF-36 might be adapted to estimate a quality adjustment value.

The remainder of the thesis (Chapters 5 to 10) reports on the empirical research. Chapter 5 presents the detailed methods of the research and the reasoning behind the key methodological decisions. The chapter begins by describing the process of adapting the SF-36, the reasons for the decisions taken by a multidisciplinary team, and the resulting classification. This is followed by a detailed discussion of the survey undertaken to obtain values for the SF-36 and the methods for estimating the preference weights using econometric methods.

The results of the valuation survey are reported in Chapter 6. These include a detailed examination of the completeness, reliability, and consistency of the data. The results of the survey are presented in the form of descriptive statistics and distributions. The remainder of the chapter considers the implications of the skewed distributions of the valuations for subsequent analyses, including the econometric modelling for estimating preference weights.

For reasons set out in this thesis, the two preference elicitation techniques used in the valuation survey were visual analogue scale rating and standard gamble. There has been an important debate about the use of the simpler VAS technique to value health descriptions and the appropriateness of transforming its results into standard gamble values using a statistical model. In Chapter 7, such transformations are estimated from the data collected in this survey and the results used to examine the validity of this approach.

Chapter 8 presents the main econometric analyses of the valuation data to estimate the algorithm for deriving preference weights for the SF-36. It includes a detailed discussion of the dependent variable, model specification, and estimation techniques (including the application of multilevel modelling techniques). Each model is rigorously examined in terms of its performance against the standard econometric tests.

Chapter 9 examines the extent to which the adaptation of the SF-36 for valuation using the preference elicitation techniques and the further simplifications brought about by the econometric modelling have reduced or eliminated the extra sensitivity of the original instrument. This is one of the ways of judging the success of this research. The criteria used in Chapter 3 to review the five preference-based measures are put to use on this new measure. The algorithms for deriving quality values are applied to five patient data sets. The quality values are also used to perform a cost per QALY analysis of alternative procedures for hernia repair from SF-36 data.

The final chapter is concerned with highlighting the contributions of the research to the theory, methodology and application of economic evaluation to health care and identifying the need for further research.

# Chapter 2

## Justification of the QALY measure

It is generally agreed in economics that it is not possible directly to measure utility in cardinal units (Deaton and Muellbauer, 1980). Economists have instead attempted to measure utility indirectly using monetary units and have developed the concepts of willingness to pay or willingness to accept in order to value intangible benefits. This approach has been widely used in transport and environmental economics but has been less popular in health economics. The QALY has been developed in health economics as an alternative measure of benefit to monetary units. A development unique to the sub-discipline. This chapter reviews the justification for using the QALY approach to value health care benefits.

The chapter is divided into seven sections. The first section defines two important concepts used in this thesis: health and utility. It is important to clarify these terms since a large part of this thesis is concerned with the relationship between them. Section two reviews the use of conventional monetary methods for benefit valuation, including revealed and stated preference methods, and the problems of using them in valuing the benefits of health care. The third section introduces the reader to the QALY, what it implies for the nature of preferences over health and the different approaches to deriving QALY values. This is not intended to be a critical review, but a simple description of the key features of the basic QALY. Sections four and five critically review these features. Section four examines the case for focusing on the health attributes of a person's utility function in health care to the neglect of its other arguments. Section five reviews the arguments for and against the other restrictions to the form of the utility function implied by the QALY. This section also examines the relative merits of two alternative measures of health, the Healthy Year Equivalents and the *ex ante* QALY (based on health scenarios), and why the QALY continues to be preferred by many health economists for measuring benefits in economic evaluation. Section six examines the use of the QALY measure to inform public decisions, and why health is afforded a special status in publicly-funded health care systems. It examines the economic case for

government intervention in health care, and the 'extra welfare' arguments that have been put forward by Culyer and others in defence of QALYs. The chapter concludes by arguing that there continues to be a strong case for using the QALY as a measure of benefit in health care.

## 2.1 Defining health and utility

The term utility has a number of different meanings in economics (Richardson, 1994). It is used in this thesis to refer to how desirable an individual finds one commodity or characteristics of a commodity compared to another. Economists generally agree that utility cannot be measured directly on a cardinal scale (Gravelle and Rees, 1981) but it is claimed there are ways of obtaining values which reflect it using measures such as willingness to pay or the QALY.

There are also differences in meaning attributed to the term health, and in particular its scope. Some health economists prefer a narrow, medical view of health (Evans and Wolfson, 1980; Donaldson, 1993). Take for example the decision by Evans and Wolfson (1980) to: "... *follow the lead of efficiency researchers, and conceptualise health status for inclusion <u>within</u> the utility function in its narrow, negative, but more or less objectively measurable form" (P. 16).*This narrow view contrasts with the well known World Health Organisation definition of health as a *"State of complete physical, mental and social well-being and not merely the absence of disease"* (WHO, 1948). A criticism of the WHO definition has been that it is indistinguishable from the concept of utility (Evans and Wolfson, 1980). Whilst acknowledging the definition is both broad and ambitious, it has been very influential in the development of health status measures (Ware, 1987).

Health status measures over the last twenty years have encompassed multiple dimensions, and included more than simply the absence of disease and the associated clinical symptoms. These measures include those concepts which lay people themselves regard as part of their health-related quality of life (e.g. Hunt et al., 1986; Bergner et al.,1981) and hence dimensions such as role and social functioning, leisure activities,

energy and mood, as well as the more conventional domains of physical functioning and pain. It is this broader definition of health which is used in this thesis, and underlies the SF-36 (See Chapter 4).

## 2.2 Monetary measures of benefit

### 2.2.1 Theoretical basis

The amount an individual consumer is willing and able to pay was first suggested as an indicator of utility by Dupuit (1844) and was subsequently developed by Marshall (1890). A 'Marshallian' demand curve (1890) indicates the maximum a consumer is willing and able to pay for a good or service and therefore the area under the curve represents the value of a good to the consumer, known as consumer surplus.

Hicks (1939, 1941) showed that a Marshallian demand curve reflects the twin effects of a price change, namely an income effect and a substitution effect. In Hicks's reformulation the value of a change in the quantity (or price) of a good is the budget change which would restore an individual to his/her initial utility level, called the compensating variation (CV). CV represents the amount of money the individual is willing to pay in the case of a gain, and the amount he/she is willing to accept for a loss. An alternative measure is the equivalent vatriation, which is the budget change that would move the consumer to the new utility level after the change. This is the amount of money a consumer is willing to accept instead of a gain, and the amount he/she is willing to pay to avoid a loss. The CV of a price increase from $p_0$ to $p_1$ becomes the EV of a price decrease from $p_1$ to $p_o$. The consumer surplus (CS) is an approximation for these two measures for given change, but is easier to observe than either CV or EV.

CV and EV are well established in theory as measures for estimating the value to the individual of goods and services, whether or not the commodities are traded. One is not preferred to the other on theoretical grounds, since they only differ in terms of whether the initial or final state is taken as the point of reference. Both measures are regarded as

superior to CS. However, CS can be a good proxy for CV or EV, where it has been argued that for small changes it makes little practical difference (Willig,1976). Most applied work using data from market transactions measures the CS, since this, unlike theoretical compensations, can be observed (Friedman, 1984). The problem for economists has been the practical difficulties of measuring any of them.

### 2.2.2 Monetary valuation in practice

Revealed Preferences

Traditionally economists have favoured estimating monetary valuations from the preferences revealed by consumer behaviour (Kroes and Sheldon, 1988). Actual decisions are thought to provide a more valid indicator of consumer preferences than simply asking someone what he/she would do in a stated preference questionnaire. The application of the revealed preferences (RP) approach is, however, often limited by the absence of suitable data for undertaking econometric analysis. Typically, there are an insufficient number of independent observations with adequate variation in order to estimate a model across all variables of interest (Kroes and Sheldon, 1988).

RP methods are also not appropriate in the health care field for a number of more fundamental and well-documented reasons (Arrow, 1963; Culyer, 1971a; Mooney, 1986). They require the outcomes of each alternative option to be perfectly known to the consumer. In health care, the consumer is often ignorant of what alternative treatments exist, and he/she tends to have a very poor understanding of the likely effect of treatment upon their health (McGuire et al., 1988; Donaldson and Gerard, 1993). This ignorance is an important reason for seeking medical advice, since doctors are assumed to be better informed. The doctor therefore acts as the patient's agent in the consumption of health care. Furthermore, the patient cannot be sure that the doctor is acting in his/her best interests (i.e. is a perfect agent) and has taken due account of his/her preferences. To compound this problem further, the consumer rarely pays the full market price for health care at the point of consumption since he/she usually has medical insurance or there is some form of Government funding (Donaldson and

Gerrard, 1993). For these reasons, it cannot be assumed in health care that a consumer's expenditure patterns accurately reveal his/her preferences.

Stated Preferences

Difficulties with RP methods have led to the adoption of a range of techniques under the broad heading of 'stated preference' (SP) methods or contingent valuation. These methods ask respondents to express their preferences for a hypothetical set of alternatives. The consumer is asked in the case of CV, for example, the maximum he/she would be willing-to-pay (WTP) for a gain, but there is no formal transaction at the time of questioning. Researchers in health care and other fields have tended to prefer to use WTP rather than willingness to accept (see Donaldson, 1996 for a review of the arguments).

The advantages of using stated preferences rather than revealed preferences stem from the control it gives the researcher to explore the specific situations he/she wishes to value. The researcher can exclude the extraneous factors found in real life, and focus on the key variables. SP has been widely used by economists working in the areas of transport (Bates, 1988) and the environment (Swallow et al., 1992; Opaluch et al., 1993; Adomowicz et al., 1994), as well as in market research (Cattin and Wittinck, 1982). In health care, SP methods have the potential advantage of being able to obtain the preferences of the patient rather than the doctor. Donaldson (1993) was able to locate 24 studies in the health economics literature using WTP methods in health care. WTP continues to be advocated by many health economists (Tolley et al.,1994) and its use has recently been revived in the UK by Donaldson and colleagues (see for example Donaldson et al., 1995a & b, 1997). They have argued that one of its advantages is that it can be used to value all the benefits of health care interventions rather than just health, whether from the perspective of the patient considering different options for treatment or the citizen contemplating treatments for different people (interpersonal comparisons are examined below). WTP allows the benefits of health care to extend beyond simply health to such things as re-assurance from information and satisfaction with the processes of care (Donaldson, 1993; Mooney and Lange, 1993).

The obvious criticisms of SP methods stem from the problem of validating the responses of subjects to hypothetical choices. A person considering a hypothetical and often remote prospect may not be able or indeed willing to give an accurate indication of his/her actual preferences. One problem is the risk of strategic behaviour where the respondent alters his/her response in order to promote their interests. In health care, a patient might exaggerate the amount he/she claims he/she would be WTP for a treatment in order to increase the likelihood of the service being provided.

There are also concerns with the ability of SP questions to elicit preferences, even from honest respondents, for different amounts or components of a good. It has been argued that the amount a person is WTP may not be sensitive to the amount of the good being considered. It has been suggested, for example, that respondents experience a 'warm glow' effect from expressing a WTP for a public good regardless of the size of the benefit. This results in the widely observed part-whole bias problem, where the summed amounts an individual is prepared to pay for the components of a commodity exceed the amount he/she would pay for the commodity as a whole. In a telephone poll in which Toronto residents were asked to state their WTP for the preservation of fish stocks, Kahneman and Knetch (1992) found that the median WTP for 'all lakes' in the province was only slightly higher than for a small proportion of lakes. There is, however, some dispute in the literature as to whether part-whole bias is a genuine problem or simply the result of a misspecification of the good (Mitchell and Carson, 1989). To address this concern, Bateman et al. (1997) tested for the existance of part whole bias by asking students to consider how much they would pay for vouchers for parts of a meal at a restaurant compared to the whole. They found evidence for a significant part-whole bias effect.

The extent of part-whole bias in health care is unknown, but there is evidence from a number of studies to suggest it could exist. In a study of WTP for safety measures to reduce the risk of injury, the amount respondents were WTP was found to be insensitive to the size of risk reduction and to changes in the severity of injuries (Jones-Lee et al., 1993). A recent clinical trial of patients with chronic obstructive pulmonary disease (COPD) found WTP to be less sensitive to health change than scores of a health status

questionnaire (O'Brien and Viramontes, 1994). These results may also have been the consequence of the limited amount of disposable income available to many patients, particularly in an elderly and severely disabled group such as COPD patients, which gives them little scope to express strength of preference.

In the health care context, people may not be accustomed to paying the full price for medical services at the point of use. This may have implications for the validity of their responses, and lead to problems such as part-whole bias. It may also have implications for the number of useable response. Recently published studies using willingness to pay in health care have reported useable response rates return of 58% in a study of patients on a waiting list for gallbladder surgery (Donaldson and Shackley, 1997), 55% in women attending a bone mineral measurement service (Donaldson et al., 1997), and 69% amongst women booking at a Maternity Unit (Donaldson et al., 1995). These compare, for example, with rates in excess of 80% in six out of seven studies using the EQ-5D and EQ-6D alongside numerous other health measures (See review in Chapter 3) The lower rates achieved with WTP was from non-return, non- completion of returned questionnaires, and 'protest' responses. Lower response rates may reduce the representativeness of the answers and increase the sample size required to establish the significance of differences.

### 2.2.3 Interpersonal comparisons of utility

For economic evaluation it is necessary to derive aggregate values of benefit and hence make interpersonal comparisons of utility. CV and EV were developed for assessing the utility consequences for an individual of a change. The aggregation of benefits across individuals involves normative judgements. In welfare economics, the test used to determine whether a change leads to an unequivocal improvement in the welfare of society is the Paretian criterion: which is, that a change should only be regarded as an improvement if it makes at least one person a better offer without any one else being worse off. Resource allocation decisions in health care, and indeed elsewhere in public policy, typically involve comparisons of alternatives where there are losers as well as gainers. A solution to this problem was suggested by Kaldor and Hicks, who extended

the Pareto principle to allow for the possibility of the gainers compensating the losers (Kaldor 1939; Hicks 1939, 1941). This 'potential' Pareto improvement criterion implies that if the WTP by the gainers exceeds the amount the losers are willing to accept, then the change should go ahead. The compensation need not be paid, but it has been claimed that the test nonetheless permits a comparison of interpersonal utility[1]. It therefore provides an argument for using stated WTP and willingness to accept (WTA) data in an economic evaluation.

A criticism of the compensation test has been the distributional implications of using aggregate WTP. It could lead to a reallocation of resources to projects favoured by the rich, since they would be expected to have a lower marginal utility per pound than the poor (given a diminishing MU per pound), and hence be willing and able to offer more money or seek more compensation for any given change. In principle, however, it should be possible to diagnose ability to pay problems and correct for them using distributional weights (Friedman, 1984). Donaldson (1995) has advocated a more explicit approach of WTP results being broken down by income group. This will reveal whether income alters the relative benefit of different projects and to make adjustments as appropriate. These solutions are dependent on a sufficient number responding to the survey in each income group, particularly in the poorer income group where response rates tend to be lower in surveys.

There is more fundamental concern about whether it is appropriate to base social valuations of public goods on utility (Sen, 1979). Welfare economics conventionally assumes that the value of goods or services, private or public, stems from the individual's own assessment of utility. This is a value judgement whose ethical basis for making social judgements, such as resource allocation in health care, has been questioned by some economists (Sen, 1982; Culyer and Wagstaff, 1993). As will be

---

[1] Skitovsky has pointed to a paradoxical result where it is possible for a given state 'x' to be superior to state 'y' using the Kaldor-Hicks test, and yet once 'x' has been attained for it to be possible to show 'y' is superior. It arises from differences between the measures of compensation and equivalent variation, and the larger the changes associated with a course of action for individuals, the higher the chance of this ambiguity arising (Friedman, 1984). There is little in the literature on the practical relevance of this problem and none in health care. Changes in resource allocation in health care often have major implications for individuals (e.g. hip replacements, transplant surgery) and hence there must be a risk of it occurring.

argued later in this Chapter, there is a case for focusing public resource allocation decisions on the characteristics of people, such as their health, rather than on their utility.

### 2.2.4 A summary of the case against monetary measures

It is stated preferences rather than those revealed from actual decisions which are used to derive monetary valuations of the benefits of health care. This raises questions about the validity of responses to hypothetical questions: individuals may behave strategically, or be prone to problems such as part whole bias. There is also some evidence of lower responses compared to health questionnaires such as the EQ-5D. Of more concern in the health economics literature has been the distributional implications of using willingness to pay (Brooks, 1995). Finally, it could be argued that to focus money detracts from the main objective of a publicly-funded health care system of promoting health (see next section). For one or more of these reasons, many health economists have preferred to use non-monetary measures of benefit (Brooks, 1995; Culyer, 1989; Feeny and Torrance, 1989), such as the QALY.

## 2.3 The QALY

The basic QALY measure of a year in full health contains two elements. The first is a value or utility score for states of health between zero and one, where zero is death and one is full health (Torrance, 1986). The second is the length of time a person spends in each state of health. The simplest application of the measure would be a chronic condition with a single health state where the total number of QALYs is the product of the value of the state and the number of years in the state. The QALY is defined by a bivariate utility function as follows (Pliskin et al., 1980; Miyomoto and Eraker, 1985) :

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

Where Q is a chronic state of health, V(Q) is the value of that health state and T is survival in years. For a more complex and realistic lifetime profile of differing health

states, the QALY measure is calculated as the sum of the health state values U ($q_i$) at each time period i.e.

$$U(Q_T) = t \sum_{i=1}^{T} U_i (q_i) \qquad (2)$$

Where t is the period of time (usually in years or proportions of a year to calculate QALYs) in each health state ($q_i$ )

The QALY measure implies the following assumptions about the individual's utility function :

1) All non-health benefits from health care have a value of zero.
2) The time preference rate for health is zero[2].
3) The value of a health state is independent of the time spent in the state.
4) The value of a health state is independent of the health state(s) which went before it or are expected to come after it.
5) The individual is risk neutral.

These assumptions make the QALY a very versatile measure, and one which can be used in decision tree analysis and the application of markov models (Weinstein et al., 1980). However, they are very restrictive and have been criticised in the literature for misrepresenting an individual's preferences for health care (Gafni et al., 1993). (The cases for and against these assumptions are examined in the next three sections of this chapter).

There are three methods for deriving the quality adjustment or health state value for calculating QALYs. The first method assigns a group of patients to a standardised classification of health, usually by asking the patients to complete a questionnaire on their health, and a value is derived for their health state from a set of off-the-shelf

[2] It has been recommended by some health economists that QALYs are adjusted for time preference using standard discounting procedures (Williams, 1985; Gudex, 1986). This assumes a constant rate of time preference. Johannesson et al. (1993) have argued this is not valid for the QALY model.

preference weights. There are five of these preference-based measures of health (see Chapter 5). The weights would have been obtained by asking a separate group of respondents, such as another group of patients or members of the general population, to value the health states described by the classification using preference elicitation techniques. The most commonly used techniques for eliciting preferences are the visual analogue or rating scale (VAS), magnitude estimation (ME), standard gamble (SG), time trade-off (TTO) (Torrance, 1986) and person trade-off (PTO) (Nord, 1993). These are briefly described in Table 2.1 (reproduced from Torrance, 1986).

The second method is to develop bespoke descriptions or vignettes of the health states experienced by patients receiving different interventions and value these using one of the preference elicitation techniques. These vignettes can be based on interviews with patients (e.g. Cook et al., 1994). The third is to ask patients to value their own state of health. (The advantages and disadvantages of these different methods for deriving health state values are reviewed in Chapter 3. The elicitation techniques are reviewed in Chapter 5.)

## 2.4 Health as the main source of benefit from health care for the individual

### 2.4.1 Health in consumer theory

Consumer theory is concerned with predicting consumer choice between bundles of commodities given a set of axioms (Deaton and Muellbauer, 1980). It assumes the consumer has perfect information about all commodities in his/her choice set, including health care, and seeks to maximise utility subject to his/her budget constraint. The utility function of such a person, individual A, can be expressed as:

$$U_A = U_A (X_{1A}, X_{2A})..., HC_A) \qquad (3)$$

where $X_1 \ldots X_A$ are the list of commodities available in the choice set, and health care ($HC_A$) is treated as another commodity.

Conventional theory assumes the consumer obtains utility directly from goods and services. This is a very restrictive view of consumption since it does not consider why a consumer prefers one bundle of goods to another, except in terms of 'taste'. Lancaster's important contribution to theory was to argue that we consume commodities for their characteristics (Lancaster, 1966, 1971). As Lancaster (1966) argues: *"Utility or preference orderings are assumed to rank collections of characteristics and only to rank collections of goods indirectly through the characteristics that they possess"* (p. 133). Many goods will have multiple characteristics, thus for example a car has speed, comfort, aesthetic properties, sex appeal and so forth. These are the determinants of the value of a car to a consumer. Similarly, it could be argued health care is not consumed for its own sake, but for its attributes. The process of consuming health care can be extremely unpleasant, such as staying on a hospital ward, or having an invasive diagnostic test, and plainly these are not desirable activities in their own right. The patient consumes these health services for the expected benefits they will bring in terms of better health.

Lancaster was able to show how the characteristics of goods approach can be incorporated into conventional consumer theory by assuming consumers seek to maximise a utility function with the characteristics of commodities as its arguments. The budget constraint continues to be expressed in terms of commodities, but in order to provide a link with the utility function, a second set of constraints is added to represent the 'consumption technology' of the relationship between characteristics and commodities:

$$c_j = c_j (X_1,.., X_n, h_c) \qquad (4)$$

where $c_j$ is a vector of characteristics of goods, including health care ($h_c$).

Health has been seen by some health economists as a 'Lancastrian' characteristic of health care, as it is for many other commodities, such as seat belts, fresh fruit and in a negative sense, smoking (McGuire et al., 1988; Ryan 1992b). However, this is an oversimplification since as Wolfson and Evans (1980) have argued: *"This (health*

*status)..is not just a particular case of the Lancastrian or characteristics approach to consumer behaviour. Health status is not a characteristic of a commodity, but of a person (or group)"* (first parentheses by author) (p.14). This may seem a little pedantic, but it is an important distinction. Health status cannot be traded either directly as a commodity or indirectly as a characteristic of a commodity. Health is a characteristic of a person. Furthermore, as Sen (1979) pointed out, the characteristics of a good or service on their own do not tell us what the good or service could do for the individual person since this will depend on the characteristics of the person.

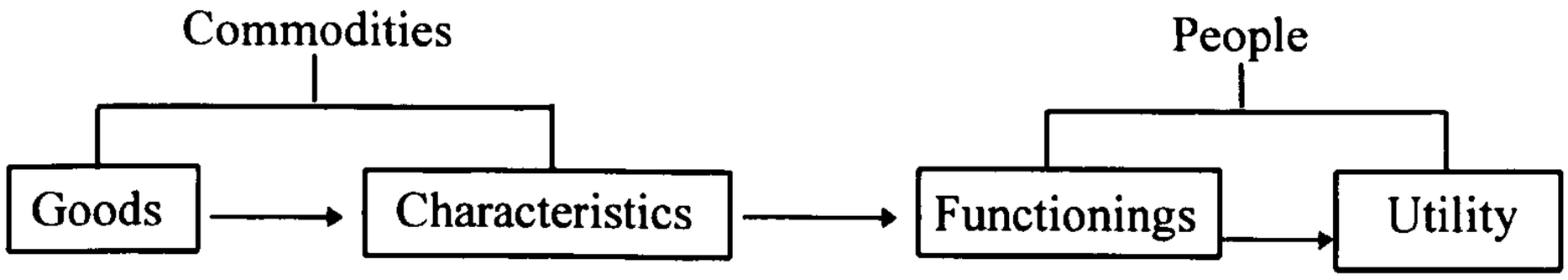
Becker's household production approach to consumption has more direct relevance to understanding the relationship between health and health care (Becker, 1962). Grossman applied Becker's model to the demand for health by making health a characteristic or 'fundamental commodity', but also distinguishing between investment as well as consumption benefits. The investment model treats health status as a durable capital stock yielding healthy days for the individual (Grossman, 1972). The individual is then assumed to maximise a life-time utility function which includes healthy days as an argument in its own right, and via also its impact on earnings.

Grossman's model has attracted considerable attention in the health economics literature (Wagstaff, 1991). It has been important because it has demonstrated how the demand for health care is derived from the demand for health. It also recognises health is a characteristic of people rather than of commodities. Health status has thus been incorporated into economic theory, and is no longer seen solely as the province of other disciplines, such as epidemiology. However, the model can be criticised for the unreality of its assumptions about the availability of information to consumers regarding the marginal efficiencies of current and future investment in health via different health care inputs (McGuire et al., 1988).

Another theoretical development looking at the reasons for consuming commodities was put forward by Sen (1979). He has proposed a chain running from goods to utility (Figure 2.1). On the left hand side, the sequence starts with goods (e.g. a bicycle) followed by Lancaster's notion of the characteristics of goods (e.g. transportation).

Consumer theory would conventionally proceed from here straight to utility, but Sen has added the intervening category of functionings (such as the ability to move in the bicycle example), which are characteristics of people. Culyer (1989a) has constructed a chain running from health care to utility with attributes of health as the person's characteristics (Figure 2.1).

**Figure 2.1: The chain from goods to utility**




| | | | |
|---|---|---|---|
| Bicycle | Transportation | Ability to move | |
| Utility | | | |
| Health Care | Timeliness, clinical efficiency, readmission rate etc. | Being reassured, being able to feed oneself, able to earn one's living. | Utility |

*Sources: Sen (1979), Culyer (1989a).*

Sen's model can be specified more formally in the partial equilibrium space of the function "good health". Pereira (1989) has specified it as follows: $X_A$ is the vector of commodities possessed by individual A, C(*) is a function for converting $X_A$ into characteristics, $f_A$ (*) is a production possibility function transforming characteristics into functionings, and $F_A$ the full set of production functions $f_A$. Any one of the production functions may be chosen by individual A and this is what Sen calls the capabilities set. A vector of health states can therefore be given by $h_A$, where:

$$h_A = f_A (C (X_A)) \qquad (5)$$

The well-being of person A can be seen as a value function containing the vector $h_A$, as follows:

$$V_A = V_A (f_A (C (X_A))) \qquad (6)$$

The value of health status $h_A$ could be measured in terms of QALYs. Buckingham (1993, 1995) has suggested that QALYs can be justified in terms of a utility function which includes years in full health as a numeraire:

$$U_A = f (H_A, L_A, P_A, W_A) \qquad (7)$$

where the utility of individual A is a function of his/her health status ($H_A$), length of life (L) or probability of death (P), and expenditure on other goods or characteristics (W). A WTP approach to valuing health would involve asking people the change in W required to compensate them for a change in H. Buckingham (1993, 1995) and Richardson (1994) have argued that other arguments of this function could be used as the numeraire, and in the context of health care, length of life might be more appropriate than W. They point out that all numeraires have problems as measures of strength of preference: wealth will be 'contaminated' by ability to pay, length of life by time preference and probability of immediate death by attitudes to risk. There is, they claim, no reason *a priori* for preferring one to the other.

### 2.4.2 Health as the main benefit of health care

The focus in health economics has been on the health benefits of health care to the exclusion of non-health benefits. This position has attracted criticism for being too narrow as a representation of individual utility functions (Mooney, 1994; Donaldson, 1993 and Ryan, 1992a). Ryan (1992a), for example, cited the work of a sociologist (Fitzpatrick, 1991) who found the following to be important to patient satisfaction in health services: *"humanness, informativeness, overall quality, competence, bureaucracy, illness, cost, facilities, outcome, continuity and attention to psychological problems"* (p.8). Donaldson (1993) has highlighted such things as dignity and autonomy in relation to long-term care and community services.

The importance of excluding non-health outcomes partly depends on how health is defined. Physical limitations and pain from, for example, surgery, would be included in most definitions of health. As would the psychological consequences of surgery (see review in Chapter 3). Even humanness, autonomy and dignity have been included in some measures of health (Wilkin et al., 1992) though not the preference-based measures of health. However, some of the benefits found by Fitzpatrick (1991) may not be incorporated into a measure of health.

The exclusion of non-health benefits from the measure intended to reflect individual preferences can only be justified empirically. Existing evidence suggests that non-health benefits can be valued. Ryan (1992b) has found that people value the benefits of invitro fertilisation, for example, despite the lack of any direct health benefits. While Mooney and Lange (1993) found evidence from a WTP study that women value the information about an hereditary renal disorder, or whether or not they were going to terminate their pregnancy. However, these two studies were looking at comparatively peripheral activities and not the core services provided by most health authorities. Specifically, they have not addressed whether individuals would be willing to sacrifice health status in order to obtain, for example, a more convenient or friendly service. This is an important issue that requires empirical research. Should the evidence suggest that individuals are prepared to make such a trade-off, then another solution would be to value the non-health benefits in terms of QALYs.

## 2.5 The theoretical basis of the QALY model

It has been argued by some economists that the theoretical foundation for using years in full health to measure utility lies in expected utility theory (EUT) and this has resulted in a technical literature setting out the axiomatic basis for the QALY model of individual preferences (Pliskin et al., 1980; Miyamoto and Eraker; 1985).

### 2.5.1 Expected utility theory and the axiomatic basis of the QALY model

It was originally Bernoulli (1738) in the eighteenth century who argued people seek to maximise their expected utility from uncertain prospects rather than the expected value. It took more then two centuries before the axiomatic basis of this theorem was first developed by Von Neumann and Morgenstern (1944). They were able to show how an individual obeying a set of axioms would be an expected utility maximiser, and subsequent analysis has led to the identification of three essential axioms: ordering (completeness, transitivity and reflexivity over all prospects), continuity and independence (Darnell, 1992). As Machina (1982) comments in a review of expected utility theory (EUT), "It became generally recognised that expected utility theory depended crucially on the empirical validity of the so-called 'independence' axiom" (P.277-278). This axiom implies an individual's utility function is linear in the probabilities of the outcomes (i.e. $EU = \Sigma U(X_i) P_i$ where $\Sigma P_i = 1$).

The attraction of EUT in the health care context is that most medical decisions involve uncertainty. Even the most routine of interventions have a risk of negative outcomes, with fatal outcomes in the case of surgery (O'Brien, 1986). The independence axiom of EUT is extremely useful analytically since it is able to break down complex medical decision problems into manageable components. It allows each outcome to be valued separately, and then an expected utility to be calculated by summing the products of each outcome and their probability, since the outcomes are assumed to be independent of one another. This avoids having to ask patients to value an entire prognosis of outcomes and probabilities at once. A further advantage is the applicability of the health state valuations to other interventions.

The link between EUT and the QALY model was first made by Pliskin et al., (1980) who identified three additional conditions which must hold for the QALY to be a valid cardinal measure of utility: mutual utility independence between life years and health status (assumed to be single dimensional), a constant proportional trade-off of life years for health status, and a constant risk attitude. Mutual utility independence exists between Life Years T and Health Status Q if preferences for lotteries involving T, with

Q held constant at $Q_0$, do not depend on the level of $Q_0$, and lotteries over Q are independent of the fixed level $T_o$. Therefore an individual who is indifferent between 10 years in full health Q* and a lottery with a 50:50 chance of 4 or 16 years in Q*, would also be indifferent with Q* replaced by an ill-health state $Q_i$.[3]. Constant proportional trade-off holds if the proportion of remaining life years an individual is willing to sacrifice for an improvement in health status from any state $Q_1$ to any other level $Q_2$ does not depend on his/her absolute number of remaining life years[4]. For example, an individual who was indifferent between 5 years in Q* and 7 years in state $Q_i$ would also be indifferent between 20 years in Q* and 28 years in $Q_i$.

Non-neutral risk attitudes can be incorporated into the QALY model in the following way (Miyamoto and Eraker, 1985):

$$U(Q, T) = [V(Q) \times T]^r \qquad (8)$$

V(Q) is a value function measuring the desirability of state Q. According to this model, the difference between the value of a health state and its utility is a person's constant attitude to risk represented r, where: r = 1 implies risk neutrality, r<1 risk aversity, and r>1 risk seeking. Johannesson (1994) has suggested a second specification based for the utility value of health state Q:

$$U(Q, T) = U(Q) \times T^r \qquad (9)$$

Here the risk parameter is only applied to T, since U(Q) is a utility value assumed to be equal to $V(Q)^r$. V (Q) is a proportion of healthy years and U (Q) a proportion of the utility of healthy years[5]. Miyamoto and Eraker have shown how r can be estimated by ordinary least square analysis from certain equivalent questions (i.e. asking the number

---

[3] This is important for the SG technique, since it ensures the utility index (i.e. the indifference probability) is independent of the time period given in the valuation task (Johannesson, 1995).

[4] The values for $h_i$ obtained by TTO are thereby independent of the duration specified in the task (Johannesson, 1995).

[5] As will be explained in Chapter 3, U (Q) is measured by *SG* and V (Q) by TTO or VAS.

of certain years in full health considered equivalent to a gamble involving full health (1) and death (0)).

The assumptions of the risk adjusted QALY model considerably simplify the empirical task of evaluating the benefit of different treatments (Weinstein et al., 1980). It has been extremely useful in empirical research because U(Q), the utility of a given health state, and the person's constant attitude to risk (r), provide simple, generalisable measures for use in decision tree analysis.

### 2.5.2 Criticisms of the QALY model and the alternatives

The QALY model has been criticised for the restrictions it places on the relationship between health and duration, its handling of risk, and more generally evidence for the violation of the axioms of EUT.

<u>Health state values and time</u>

The QALY model has been criticised for the restrictive assumptions it places on the relationship between health status and time in the utility function. It implies a special case where the utility function is linear and separable over time. The value of a health state is assumed to be independent of the time spent in the state, when it occurs (i.e. time preference) and in what sequence of states it occurs. It has been recommended by some health economists that time preference be incorporated by standard discounting procedures (Williams, 1985; Gudex, 1986). This assumes a constant rate of time preference (d), and results in a revised QALY model:

$$U(Q_T) = \sum_{i=1}^{i=T} \frac{1}{(1+d)} U_i(q_i) \qquad (6b)$$

There is evidence, however, to suggest the assumptions of both the standard QALY model and the discounted version are violated. Sackett and Torrance (1978), for example, asked patients and members of the general population to value a variety of health states, including hospital dialysis, for durations of three months, eight years and

the rest of their lives, and found the mean daily health state utilities declined with duration. These results suggest it might be necessary to estimate separate utility values for health states over different durations. More generally Richardson and colleagues (1990) have argued that the utility of a health state may be directly related to a person's prognosis:" *A poor health state may be more tolerable if it is perceived as a temporary hardship to be endured to obtain subsequent health. Conversely, the enjoyment of an otherwise satisfactory health state may be diminished by the knowledge that it will end in suffering and death".* (P.15).

It is equally plausible that a person learns to adjust to a health problem and this reduces its impact on their quality of life. The practical importance of these violations is not known.

Risk attitude

There is some evidence to suggest that the mean attitude of patients to risk in the health care context is close to neutrality. A study by Miyamoto and Eraker (1985) of 46 patients with symptomatic coronary artery disease obtained a geometric mean value of r of 1.03, where one indicates risk neutrality. In a recent unpublished study of 163 women with early stage breast cancer, the arithmetic mean r was 1.18 (Shiell et al., 1995). In both studies, attitudes to risk varied enormously between respondents. At the individual level, the values of r ranged from 0.22 to 12.95 and 0.31 to 6.46 in the two studies respectively. Furthermore, the reliability of these estimates at the individual level can be questioned given the low number of questions used to estimate r (i.e. three and four respectively in the studies mentioned).

Loomes and McKenzie (1989) found evidence from empirical research into risky prospects involving wealth that does not lend support to a constant risk attitude. In health, there is also some preliminary evidence to suggest it may not hold. In another study by Miyamoto and Eraker (1989) of 44 undergraduates, it was found the same hypothesis of constant risk attitude over survival was violated for a substantial proportion of subjects.

Violations of the axioms of EUT

EUT has been pre-eminent in the field of individual decision-making under uncertainty since the Second World War (Schoemaker, 1982). A feature of EUT is that it generates "bold and testable predictions" (Appleby and Stammer, 1987) and this has led to a considerable amount of empirical work (for a survey see Appleby and Stammer, 1987). Given the role of EUT in QALY measurement it is important to review this work and its implications for QALYs.

There is a considerable body of evidence in relation to prospects involving wealth, and an increasing amount of evidence in the field of health, which raise serious doubts about the descriptive validity of the restrictions imposed by EUT (Machina, 1987; Loomes and McKenzie, 1989; Loomes, 1993). The earliest example of this was the well-known Allais Paradox, where a change in the common consequences of two gambles was found to lead to a reversal of preferences for most subjects; a violation which could not be explained by EUT (Allais, 1979). Some economists argued this was merely an aberration, and when the inconsistency was explained most individuals would conform to the independence axiom (Savage, 1954; Ellsberg, 1961). This has happened after the extensive discussion with subjects in one study (MacGrimmon, 1968), but in more neutral discussions conformity with EUT was not achieved (Slovic and Tversky, 1974). There are now many other examples of EUT violations, including a common ratio effect (where the absolute but not relative probabilities are varied), isolation effect (i.e. where 'accumulators' do not simply 'boil down'), and finally a reflection effect (the preference reversal over gains has a mirror image for losses, thus the way in which gambles are framed influences the results). There is even evidence of non-transitive patterns of preferences (Lindman, 1971; Lichtenstein and Slovic, 1973). Although this violation has been shown in laboratory experiments, there is little evidence of the impact on actual choices (Darnell, 1992).

It is difficult to gauge the importance of the violations of EUT for the QALY measure. There has been no empirical work to assess their impact on choices in the health care context. Furthermore, there are those who do not believe the validity of the measure

depends on its alleged basis in EUT (Buckingham 1993, 1995; Richardson 1994). These issues are taken up later in section 2.5.

### 2.5.3 Alternative measures

Healthy Year Equivalents

To overcome the shortcomings of the QALY model, Mehrez and Gafni (1991) propose a utility function in health which does not constrain the relationship between quality and quantity. It is a measure, they claim, which 'truly' reflects a person's utility function over quantity and quality of life while retaining the intuitive appeal of a year in full health. To distinguish it from the QALY, they have named their new measure the Healthy Year Equivalent (HYE). In the case of a chronic health state, an HYE (H*) is defined as follows:

$$U(\bar{Q}, H^*) = U(Q, T) \qquad (10)$$

H* is the number of years in perfect health ($\bar{Q}$) such that an individual would regard it as equivalent to T years in state Q, where $H^* < T$, $Q < \bar{Q}$, and $\bar{Q}$ is set to one.

The more general case is a lifetime profile of health states, vector Q = [Qi], where qi is the individual's health state at the ith period (measured in years, though smaller units of time could be used). The HYE is defined as: find H* such that

$$U(\bar{Q}_{H^*}) = U(Q_T) \qquad (11)$$

The generality of the HYE model can be seen from this last equation. The HYE represents a measure of an entire lifetime scenario, and therefore does not impose an additional assumption about an individual's attitude to time, or the relationship between quality and quantity. Mehrez and Gafni claim it represents a Von Neumann Morgenstern (VNM) utility function.

In order to obtain the HYE for a health profile, Mehrez and Gafni (1991) have proposed a two-stage procedure using the SG technique for eliciting preferences (see Chapter 3 for review). In the first stage, the respondent is asked to consider the lifetime health profile, say the chronic state of (Q, T), and an alternative of full health (Q) with probability P and a 1-P chance of immediate death ($Q^D$). The respondent undertakes a conventional probability equivalence gamble and hence to determine P* such that:

$$(Q, T) \sim P^* (\bar{Q}, T) + (1-P^*) (Q^D, T) \qquad (8)$$

Stage two is a certainty equivalence question where the choice is between the right hand side of equation (8) and years in full health. The respondent is asked how many years (H*) in full health he/she would regard equivalent i.e. set H* such that:

$$(\bar{Q}, H^*) \sim P^* (\bar{Q}, T) + (1-P^*) (Q^D, T) \qquad (9)$$

The SG method is directly derived from expected utility theory (EUT). Given the axioms of EUT, 'P' is a cardinal index measuring an individual's preferences under uncertainty (Von Neumann and Morgenstern, 1944). SG is regarded by Mehrez and Gafni as the 'gold standard' amongst valuation techniques in health care because of its axiomatic basis in EUT, the classical theory of decision-making under uncertainty. It has been shown to incorporate a person's relative attitude to risk (Dyer and Sarin, 1982). This is regarded as important given that health care decisions are made under conditions of uncertainty. (These alleged advantages of SG are reviewed in Chapter 3.) Furthermore, this property is maintained for the measure H* through the two-stage procedure.

The HYE measure proposed by Mehrez and colleagues has been criticised on a number of grounds. The first has been the feasibility of the two stage procedure. Mehrez and Gafni conducted an experiment to assess the feasibility and reproducibility of the procedure. A sample of 32 graduate students was asked to complete the procedure for a chronic state (hospital dialysis for ten years followed by death) in two interviews, separated by four weeks. The interviews took seven minutes for the first test and five

minutes at re-test. They found a high degree of reproducibility by correlation (r = 0.78) and t-test (i.e. $H^*_1 = H^*_2$ was not rejected). However, there has been little other experience with the procedure. A second and more fundamental criticism has been the claim that the HYE algorithm used to value health profiles is redundant. In five separate papers, the HYE has been argued to be indistinguishable from a profile valued by Time Trade-off (Johannesson et al., 1993; Buckingham, 1993; Culyer and Wagstaff, 1993; Loomes, 1995; Morrison, 1995). The argument is as follows:

given the equivalence of the right-hand sides of equations (8) and (9), transitivity implies indifference between their left hand sides i.e.

$$(Q, T) \sim (\bar{Q}, H^*) \qquad (10)$$

This is precisely what is established by the TTO procedure (see Chapter 3). A direct TTO question asks a person to trade the length of time in an intermediate health state Q in order to achieve full health. The lesser number of years, X* is set so that:

$$(Q, T) \sim (\bar{Q}, X^*) \qquad (11)$$

Again by transitivity, the individual must be indifferent between the right hand sides of (10) and (11) i.e.

$$(\bar{Q}, X^*) = (\bar{Q}, H^*) \qquad (12)$$

Assuming strict monotonicity (or ''increasingness''), this final expression can only be true if X* = H*. As Loomes (1995) states, "In short, without imposing any particular functional form, we see that an EU maximiser will (in the absence of errors) always give exactly the same response to the direct TTO question as to the two-stage HYE procedure" (p.2). Gafni et al., (1993) argue that SG yields a utility whereas the TTO method provides a value. Loomes argues, however, that it is quite clear from Dyer and Sarin's (1982) work on this subject that *"for every value of V (.) there is a corresponding (unique) value of U (.), so that all outcomes which have the same V(.) as each other necessarily have the same U(.) as each other"* (p.4), thus if y ~ x, then V(y)

= V(x) and U(y) = U(x). Any differences between TTO and HYE values must be due to measurement error, or because people's preferences violate monotonicity and/or transitivity.

The complex two-stage procedure to estimate HYEs may be indistinguishable from TTO, but the more general model of preferences originally proposed by Mehrez and Gafni (1989) has been taken up by other health economists. This has been the valuation of whole health profiles.

Valuing health scenarios

Under this approach whole scenarios of health are valued at once using preference elicitation techniques, such as TTO or single stage SG. These scenarios include the sequence of health states and their duration and have been extended to incorporate the probability of the states occurring (e.g. Cook et al., 1994; Sculpher et al., 1993). The incorporation of uncertainty into the scenarios makes them more realistic and enables people's attitudes to the actual risks associated with the scenario to be included in the valuation. Cook et al., (1994) have called this the '*ex ante* QALY' approach and contrast it to the QALY model where the valuation of the health states is made ex post. These descriptions are also able to incorporate short temporary health states, as well as the processes of the care itself (Sculpher et al., 1993). The health scenarios have the potential of being more general than the QALY model.

The criticisms of this approach have been concerned with its feasibility. The valuation of every conceivable health, with all the states, durations and sequences presents a substantially larger valuation task than the QALY approach. Johannesson et al. (1995a) have argued that this approach is "clearly infeasible in the context of the types of decision-models currently used in outcomes research and health policy analysis, including Markov models" (P. 283). The extension of this approach is one way of reducing the number of scenarios for valuing, but the descriptions are then in danger of becoming too large and complex for respondents to value. In practice, the *ex ante* perspective has only been applied partially (Hall et al., 1992; Cook et al., 1994). The most detailed application of this approach has been by Cook et al., (1994) in a cost-

utility analysis of alternative treatments for gallstone disease. This study illustrates the strengths and weaknesses of the approach.

Cook et al., (1994) compared the QALY and *ex ante* scenario approaches. A QALY loss was calculated as the sum of the products of: the probability of a successful operation and its associated health status, the probability of a complication and the associated health status, and the probability of death (1 in 1000) and expected survival. The 'partial *ex ante*' approach retained the first two parts of the QALY calculation and only differed in the third part. The risk of mortality was incorporated into a scenario, along with the process of the operation and the stay in hospital following the operation. There were two of these *ex ante* scenarios, one for each type of operation:

*Operation Scenario (1):*
*You will have an operation. Your doctor has told you that there is a very small risk of dying (about one person in every 1,000 dies). After the operation you will return to full health straight away.*

*Operation Scenario (2):*
*You will have an operation. Your doctor has told you that there is a very small risk of dying (about one person in every 1,000 dies). After the operation you will be in hospital for one week and you will: have a dull gnawing sort of pain all of the time; feel sick and want to vomit most of the time; find coughing and moving painful; have constipation and will be given an enema; have trouble sleeping.*

Cook et al., (1994) found the valuations of these scenarios implied significantly larger losses from the operations than the QALY approach and this was important enough to alter the rankings of the surgical options compared to treatment by lithotripsy in terms of cost-effectiveness (though the method of costing proved to be more important). The cause of the difference is not clear, since the approaches differed in terms of <u>what</u> was valued as well as <u>how</u> it was valued. The scenarios included the operation and in the case of the second operation, the description suggests a very unpleasant experience for a week which is not included in the QALY calculation. Nonetheless, the result confirms

evidence from the general EUT literature that small probabilities of large losses tend to be valued more highly by individuals than would be predicted by the QALY model. This would lend support to the *ex ante* approach, assuming the purpose of the measure is to reflect individual preferences (an issue discussed in section 5).

However, the Cook et al., study demonstrated the problems with this approach. The descriptions of the outcomes of the operations were very brief and simplistic, and only loosely based on evidence. The authors acknowledged that *"the inability of an individual to process large amounts of information in a reliable and valid way makes such an analysis difficult"* (p.158). This was why they opted for a partial scenario approach. As a result the vignettes described only one outcome for a successful operation and one complication (i.e. common bile duct damage). This fails to reflect the considerable range of outcomes experienced by patients in terms of the extent of symptom relief (from deterioration through to substantial improvements) and complications (Nicholl et al., 1992). Furthermore, apart from death, all the outcomes are described in terms of certainty. The mean stay of an open cholestectomy operation was assumed to be a week, but this masks a very wide distribution. The mean stay following the less invasive laporoscopic procedure was assumed to be zero but this does not reflect experience reported elsewhere of one or two days in hospital, and again associated with a wide distribution (Majeed et al., 1996).

The doubtful validity of the scenarios would have seriously jeopardised the value of the Cook et al., study. It seems to be impossible to incorporate the richness and variability of outcomes associated with such health care interventions into scenarios. The QALY approach is better able to do this through the repeated use of preference-based measures of health in clinical trials.

The scenario approach also suffers from a degree of inflexibility in modelling using decision trees and markov procedures. In the QALY model, it is comparatively straightforward to test the sensitivity of the results to different assumptions about the probability and duration of the outcomes. The scenarios will be based on particular values for these variables and therefore to undertake a sensitivity analysis would involve

re-valuing the scenarios and hence substantially increasing respondent burden. The risk of overloading the respondent may restrict the scope of a sensitivity analysis (Sculpher et al., 1993).

In theory, the scenario approach makes less restrictive assumptions about the form of the individual's utility function in health. It can, in principle, avoid the assumptions about additive separability of health states through time and risk neutrality. There have been few applications of the approach, but experience from the Cook et al. study supports the reservations expressed by Johannesson et al. (1995a) about feasibility. To operationalise the concept it is necessary to exchange a set of known restrictions on the form of the utility function for empirical simplification and a significant degree of inflexibility. Far more experience is needed in applying the scenario approach, along with more studies comparing it to the QALY approach, before it will be possible to judge which is superior.

## 2.6 Social values and QALYs

The previous two sections were concerned with the justification for using the QALY as a measure to reflect individual preferences. The concern with individual preferences reflects the conventional view in welfare economics that any two courses of action should be compared and ranked in terms of the utilities of the individual members of society (Arrow, 1951; Debreu, 1959). The use of measures based on health such as the QALY, rather than utility, is a departure from this view. The question addressed in this section is whether it is appropriate to use QALYs to inform social decision making. This requires an investigation into the reasons for public involvement in health care.

A large proportion of health expenditure in developed countries is funded from public sources (Office for Economic Cooperation and Development, 1987). The usual explanations in the health economics literature for this level of involvement have been based on arguments about the government being more efficient and equitable at providing insurance cover for the public and dealing with externalities which are alleged to arise from the consumption of health care (Culyer 1971a & b,1989b). The next two

sub-sections consider their implications for the use of QALYs. These arguments have been described by Culyer (1989b) as welfarist arguments, since they are concerned with the consequences from the consumption of commodities for an individual's utility. The other approach, which Sen (1982) and Culyer (1989b) term as extra-welfarist, relaxes these restrictions on what information can be used to assess social states. It allows the non-goods characteristics of individuals (such as their health) to enter into judgements and has been more concerned with aiding decision makers in allocating resources. This view has been particularly important in the justification for the QALY measure (Culyer 1989a; Richardson 1994) and will be examined in sub-section three.

### 2.6.1 Public insurance

Disease and ill-health are stochastic in nature. The costs of such adverse events from the consumption of health care can be considerable. A risk averse individual in these circumstances will insure where the welfare loss of the certain premium is less than the welfare loss from the expected financial losses. This could be left to the private market to provide, but it has been shown in the economics literature that 'failures' may arise (Arrow, 1963) [6]. A system of public finance may result in lower costs by reducing x-inefficiencies from monopoly providers and the excess consumption for those who insure, and increase welfare gains by extending coverage (Culyer 1989b). There could also be equity advantages and these are examined in section 2.6.3.

This welfarist argument is based upon the aim of maximising individuals utility from their own consumption of insurance. There is no reason to suppose that people would only be concerned with the health benefits from the health care they received through public insurance. The question is whether this is a reasonable simplification. There are

---

[6] There are economies of scale in the provision of insurance and therefore there is a tendency to move towards a few providers or even to one. This could lead to X-inefficiencies. It is argued that private insurance markets would also be inefficient owing to a problem known as moral hazard. This arises from the fact that at the point of use insurance substantially reduces the costs of consumption to the patient, and therefore he/she is likely to consume above the socially optimum level. It would also lead to increased premiums and this would discourage some risk averse individuals from insuring. The market response to this problem has been to introduce charges, and hence reduce the level of cover, in order to discourage over-consumption. Cover would be further reduced by the problem of adverse selection. Premiums are based on the average risk for the group. It is argued in the literature that individuals will have better information about their risks, and those with low risks are likely to self-insure. This also has the effect of increasing the premiums for the high-risk groups, many of whom would be poor and may not be able to afford to remain insured.

two empirical issues to consider. The first is whether the aspects of medical care covered by insurance, presumably the more expensive and unpredictable items, are more likely to be concerned exclusively with health. The second is the problem which was addressed in section 2.4, which is whether individuals would be prepared to sacrifice health status for non-health benefit, such as accessibility, friendliness and dignity. We currently do not know the answers to these questions.

### 2.6.2 Externalities

An important argument for government intervention is the alleged existence of externalities. The earliest explanation was the direct physical externality of the risk to an individual of catching a communicable disease from another person (Weisbrod, 1961). For two individuals, A and B, this externality can be represented in A's utility function as follows:

$$U_A = U_A (X_{A1}, X_{A2},..,HC_A, HS_A (HC_A, HS_B (HC_B, HS_A))) \qquad (12)$$

where individual A's utility is a function of his/her consumption of health care ($HC_A$), other commodities ($X_{iA}$) and his/her own health status ($HS_A$). This last element is dependent on the health of individual B ($HS_B$ ) owing to the risk of catching the disease from A. The health status of B is in turn dependent on the health status of A. This 'physical externality' argument can explain why a measure of health status would be a good surrogate indicator of preferences. The argument for public involvement then rests on the free rider argument. The private market would not ensure optimal consumption patterns (i.e. where social marginal cost equals social marginal value) since the benefits derived from the consumption of health care are not restricted to the parties of the transaction. Individual A benefits more if someone else makes the donation to B's health care, since he then avoids the cost but obtains the same benefit. This results in an under-consumption of health care. One solution is Government intervention through subsidy or public provision (such as the National Health Service). In the case of the 'physical externality', the focus on health status would seem to be justified. However,

spending on preventing communicable disease accounts for only a small proportion of the public funding of health care.

Health economists have extended the externality argument to include the consumption of medical services (Pauly, 1971; Lindsay, 1969). Culyer was able to show how this 'caring' externality for health care supports a 'philanthropy-in-kind' system or a tax-subsidy over income transfer (Culyer 1971b). It implied complex variations in subsidy, depending on the income and tastes of the recipient, and thus was impractical. More importantly it does not explain why society should choose to subsidise health care rather than other goods and services. It was Culyer and Simpson (1980) who, among others, recommended incorporating the health status of others as the source of the externality, rather than their use (or lack of use) of health care (which now becomes instrumental in the utility function). In their interpretation of this 'caring' externality, Evans and Wolfson (1980) suggested the following utility function:

$$U_A = U_A\,(X_{A1}, X_{A2}, HC_A, HS_A(HC_A, HS_B\,(HC_B, HS_A)), HS_B(HC_B, HS_A), HC_B). \qquad (13)$$

The health status of B enters the utility function in its own right, as well as via its impact on A's health status. The health care received by B now enters the utility function of A indirectly via health status, thus:

$$\frac{d\,U_A}{dHC_B} = \frac{\delta U_A}{\delta HC_B} + \frac{\delta U_A}{\delta HS_B} \cdot \frac{\delta HS_B}{\delta HC_B} \qquad (14)$$

The indirect effect of health care on utility via health status is assumed to be positive, but Evans and Wolfson make no assumption about its direct influence. They argued it was probably negative because A would be benevolent, and sympathise with the negative outcome of the process of care. Individual A would usually regard any consumption of HC which does not increase $HS_B$ as 'unnecessary' and given its cost, undesirable. There is no room for the non-health benefits here. In justification, Evans and Wolfson point to the considerable concern in publicly-funded systems with ensuring

the effectiveness of health care: *"We characterise A as benevolent, wishing B <u>well</u>, but not necessarily happy."*[7]

The case for Government involvement in the face of these 'caring externalities' again rests on the free rider argument. The consequences for health policy of the externalities arising from the health status of others are, according to Culyer and Simpson (1980): *"a) measurement of health b) estimation of health production functions (...) c) determination of efficient spending patterns''* (1980, p. 228). This argument provides an important justification for the development and application of the QALY methodology since they are a measure of health, can be used to quantify the production of health, and can compare spending patterns in terms of health gain.

### 2.6.3 Extra welfare arguments

The basis of the externality argument, and in particular the free rider theorem, has been challenged in the economics literature (Sugden 1980). Sugden (1980,1982) has shown how the free rider theorem on its own "*has implications that are paradoxical, implausible and inconsistent with empirical evidence*" (P. 350). For example, under the assumptions of this theorem (and the known median income elasticity of charitable giving), an increase in the contribution of an individual to charity should <u>reduce</u> the amount other donors give by almost the same amount. For every additional £1000 from Government, there should be an almost equivalent reduction in private resources. These predictions have not been observed (Sugden, 1982).

---

[7] The arguments for a broader view of utility rather than health would imply a purely altruistic concern by A about B, and can be represented by the following:

$$U_A = U_A\,(W_A\,(X_A, HC_A), W_B\,(X_B, HC_B))$$

Here A derives utility from B's use of health care via its effect on B's utility. Her/his utility function has two components, a selfish and an altruistic part, where $W_A$ and $W_B$ are the welfare each consumer derives from his/her own consumption. The weighting attached to each component will be a matter of taste. Such a formulation implies a redistribution of wealth, but does not provide any justification for subsidising health care. Respecting the individual's choice must mean allowing him/her to choose not to allocate resources to health care (and even purchase health hazardous commodities such as tobacco and alcohol).

There has therefore been a search for alternative explanations for public involvement and these have centred around notions of equity or justice, fairness, and duty (Mooney, 1986; Culyer, 1989a & b; Pereira, 1989)[8]. Definitions of equity in health care have tended to focus on equality of access to health care or choice sets, or the outcome of health itself (Mooney, 1986). Essentially the debate is between those who seek to emphasise the role of consumer preferences and hence support notions of equal access to health care, and those who believe the focus should be on health (Mooney et al., 1991; Culyer and Wagstaff, 1993). The argument for the latter has been put by Culyer and Wagstaff (1993) who asked: *"Why should health care be a concern for equity purposes in the first place?"*.

In welfare economics, it is only the utility of individuals which is relevant in choosing between states of the world. The source of a person's utility, or the human desire being satisfied is ignored. Sen (1980; 1983) argues that focusing on utility is profoundly inadequate. This can be illustrated by examples such as torture, discrimination, and the suppression of liberty, where we do not evaluate the pleasure which may be derived from these pursuits using the same calculus as the satisfaction gained from relieving hunger or cold, or the curing of illness. The utility individuals gain from good health is held in high regard by society. Sen proposes the notion of 'basic capabilities' or 'a person being able to do certain basic things', such as being able to walk about. These might be highly regarded because they affect a person's ability to live his/her life to the full. Culyer and Wagstaff (1993) cite the philosophical literature where:*" it is argued that entities such as 'good health' are necessary for an individual to 'flourish' as a human being. Insofar as health care is necessary to 'good health', this provides a strong ethical justification for being concerned with the distribution of health care and not with the distribution of, say automobile spares, and for using the word 'need' in the context of health care and not in the context of, say, skiing holidays"* (P.452).

---

[8] Mooney (1986) and Culyer (1989b) have also explored a theory put forward by Margolis (1980) that individuals obtain satisfaction from contributing to a group. This avoids the free rider explanation and provides an interesting, if under-developed, theory for private philanthropy. However, the private act of giving and the utility gained from it, seem entirely different to contribing by taxation.

This departure from welfare economics has been advocated by a number of health economists (Culyer 1989a; Richardson 1994). It can be seen as part of the decision-aiding tradition in economics which places less emphasis on conventional welfare economics, and more on the need to help decision makers allocate resources in a way which maximises their objectives (Sugden and Williams, 1978). This approach implies that a measure should have a clear meaning to decision makers and conform to their objectives (Richardson, 1994). This tradition is not well received by many economists (see recent review by Johannesson et al., 1996), but it has been important and influential in health economics.

Extra-welfarism and the focus on health rather than utility, does not imply the abandonment of economic theory. Economics is better equipped than other disciplines to tackle what Sen (1980) has described as "*the problem of indexing basic capability bundles*". Judgements must be made, but by whom and how? Culyer (1989a) has argued that there is a role for consumer theory since it provides important insights into properties of measurement, the importance of value judgements and a set of experimental techniques for studying those values. The QALY approach is exactly this, namely the application of ideas from consumer theory to the valuation of health. However, there are two differences from the traditional application of consumer theory. Firstly, the arguments of the function are being restricted to those personal characteristics regarded by society as important, such as good health. Secondly, the values may not be those of the user, but the tax payer, the electorate or whoever is deemed appropriate. The choice of constituency is an important value judgement (Mooney, 1994).

## 2.7 Conclusion

This chapter reviewed the arguments for and against using QALYs to measure the benefits of health care. It examined the reasons why many health economists have preferred to use non-monetary over the more conventional monetary measures of benefit (Brooks, 1995; Culyer, 1989a; Feeny and Torrance, 1989).

The main non-monetary alternative has been the QALY, which provides a convenient analytical framework for measuring the benefits of health care. It does this by restricting the source of the benefits and the nature of the individual's utility function. Health is an important source of utility, and it is argued in this chapter that limiting the benefits (of health care) to health is justified since it is the main source of benefit at the individual and societal level.

The QALY measure has been shown to depend on a set of assumptions about the relationship between the value of health states, time and uncertainty. These assumptions provide the QALY with the flexibilty to be used in decision tree analysis and Markov modelling using data collected prospectively in clinical trials. There are doubts about the empirical validity of these assumptions, but there is little evidence regarding the significance of the violations for decision making. The alternative measures of the HYE and '*ex ante* QALYs' are more difficult and inflexible to apply, and exchange the set of unfounded theoretical QALY assumptions for a more complex and lengthy set of valuation tasks. The valuation tasks may prove to be infeasible in many circumstances. For these reasons, a recent expert panel appointed by the US Public Health Service concluded that QALYs continue to be the preferred measure for assessing the cost-effectiveness of health care interventions (Gold et al., 1996).

**Table 2.1: Value elicitation techniques**

**Visual analogue scale**

*"A typical rating scale consists of a line on a page with clearly defined endpoints. The most preferred health state is placed at one end of the line and the least preferred at the other end. The remaining health states are placed on the line between these two, in order of their preference, and such that the intervals or spacing between the placements correspond to the difference in preference as perceived by the subject"* (Torrance, 1986, P.18).

**Magnitude estimation**

*"Here the subjects were asked to provide the ratio of undesirability of pairs of health states - for example, is one state two times worse, three times worse etc. compared to the other state? Then, if state B is judged to be x times worse than state A, the undesirability (disutility) of state B is x times as great as that of state A. By asking a series of questions all states can be related to each other on the undesirability scale"* (Torrance, 1986, P.25).

**Standard gamble**

*"The subject is offered two alternatives. Alternative 1 is a treatment with two possible outcomes: either the patient is returned to normal health and lives for an additional t years (probability P), or the patient dies immediately (probability 1-P). Alternative 2 has the certain outcome of chronic state i for life (t years). Probability P is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is simply P; that is hi = P* (Torrance, 1986, P.20).

**Time trade-off**

*"The subject is offered two alternatives - alternative 1: state i for time t (life expectancy of an individual with the chronic condition) followed by death; and alternative 2: healthy for time x<t followed by death. Time x is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is given by hi* = (Torrance, 1986, P.23).

**Person trade-off**

*"If there are x people in adverse health situation A and y people in adverse health situation B, and if you can only help (cure) one group (for example, due to limited time or limited resources), which group would you choose to help?". One of the numbers x or y can then be varied until the subject finds the two groups equivalent in terms of needing or deserving help. If x and y are the equivalent numbers as judged by the subject, the undesirability (desirability) of condition B is x/y times as great as that of condition A. By asking a series of such questions all conditions can be related to each other on the undesirability scale"* (Torrance, 1986, P. 25).

# Chapter 3

## A review of preference-based health measures

There are currently five generic preference-based health-related quality of life measures designed to estimate the quality adjustment value used to calculate QALYs: the Quality of Well-Being scale, Rosser's disability/distress classification, the Health Utility Index versions one, two and three (HUI-I, HUI-II and HUI-III), the EQ-5D and the 15D. To understand the contribution of this research to the methodology of economic evaluation it is necessary to review these measures. Existing reviews of these measures have been restricted to applying 'psychometric' criteria (e.g. McDowell and Newell, 1987; Wilkin et al., 1992; Anderson et al., 1993) and have not taken into account the requirements of a measure for economic evaluation (Williams, 1993). This chapter reports on the first systematic review of the five preference-based measures from an economic perspective.

The chapter begins with a background section examining the reasons for the dominance of standardised preference-based measures of health for estimating health state values, such as the EQ-5D, over direct preference assessment and the use of condition-specific vignettes. This is followed by a section setting out the economic criteria for reviewing the five preference-based measures. The next section describes the search methodology used to identify the papers for the review. The remainder of the chapter is a systematic review of the five measures against the criteria. The final section considers the implications of the findings of the review for the research reported in this thesis.

### 3.1 Background

The five measures reviewed in this chapter all use the same approach to valuing health states. Patients are assigned to health state classification, usually by asking them or someone on their behalf, to complete a questionnaire. The patient's health state is valued from a set of off-the-shelf preference weights. Another approach is to develop bespoke descriptions or vignettes of the health states experienced by patients receiving different interventions and to value these using one of the preference elicitation techniques. A third is to ask patients to value their own state of health.

The debate concerning the appropriateness of using generic preference-based measures versus condition-specific descriptions is a long-standing one. There has been a concern about the relevance and sensitivity of some generic preference-based measures (e.g. Donaldson et al., 1988; Hall et al., 1992; Cook et al., 1994; Hollingworth et al.., 1995). Bespoke descriptions are likely to be more relevant to the condition. An early example was a study by Sackett and Torrance (1978) of patients with chronic renal disease being treated by hospital dialysis, home dialysis and renal transplantation. Another was a study of breast cancer screening, where Hall et al. (1992) constructed descriptions of quality of life with breast cancer. They chose not to use one of the generic measures because these measures were thought to exclude a number of aspects of life found to be important to the women themselves (e.g. diagnosis of cancer, physical appearance, certain symptoms etc.). The relevance of a generic health classification, however, depends on the condition being studied. It is designed to describe the core features of health, and for many conditions this may be adequate. For rheumatoid arthritis, for example, generic measures have been found to be as sensitive as condition-specific measures (Fitzpatrick et al., 1993).

The direct use of preference elicitation techniques on patients has the potential advantage that patients are valuing their own state of health rather than some hypothetical health state. Buckingham (1993) argues: *"To ask a person of twenty years how s/he will value health at the age of seventy is to ask an enormous amount of their imagination. To ask a seventy year old how important their health is to them is likely to result in far more valuable information"* (P. 306). The imagination required to value the generic classifications partly depends on the accuracy of the health state descriptions. A broad definition of health that takes into account the consequences for a person's work and social life, as well as physical functioning and mental health, will make it easier to imagine such states.

A disadvantage of direct utility assessment is that it has been found to be less responsive to health change than standardised health status questionnaires. In the Canadian Erythropoiten Group Study (Laupacis, 1990), statistically significant differences were

found between the experimental and placebo groups in measures of fatigue and exercise stress, and two dimensions of the Sickness Impact Profile (which in the past has been criticised for being insensitive (Wilkin et al., 1992)), but direct utility assessment using TTO did not find any significant differences. A similar result was found in a study by Katz and colleagues in a study of patients undergoing hip arthroplasty (Katz et al., 1994). Lower responsiveness means larger sample sizes will be needed in order to detect differences, resulting in a more costly trial.

In practice, the direct approach has not been widely used. (Drummond and Davies, 1991). It has encountered considerable resistance from clinical investigators concerned about the added distress to their patients from valuation questions (such as SG) that confront patients with unpalatable scenarios involving, for example, death, and resulting in patients withdrawing from a trial. It is usually more acceptable on ethical grounds to collect the descriptive data from patients in a trial, and obtain the values outside of the trial.

The generic approach has important advantages. Using the same measure across conditions ensures comparability between studies in terms of values. The other two approaches use valuations obtained from different groups of respondents in each study. Generic measures are therefore more suitable for cross-programme comparisons and for informing decisions about resource allocation between programmes. The generic approach is easier to use and has off-the-shelf values, whereas a condition-specific approach requires the descriptions to be re-constructed in each study and be valued, and direct preference elicitation will involve asking patients difficult and potentially upsetting questions. The advantage of ease of use, however, depends on the extent to which the descriptions and valuations are valid and these are reviewed in detail for each of the five preference-based measures later in this chapter.

### 3.2 Review criteria

The five preference-based measures are reviewed in terms of their practicality, reliability and validity. Practicality and reliability are reasonably uncontroversial

(Torrance, 1976; Dolan et al., 1996), if somewhat neglected criteria in the economics literature, whereas validity is a major area of disagreement. These will now be considered in turn.

### 3.2.1 Practicality

The practicality of an instrument depends on its acceptability to respondents, and the cost of administration (e.g. in terms of time). These are plainly related issues since a lengthy and costly instrument is likely to be unacceptable to many respondents and hence may prove to be infeasible. Acceptability is a function of length and complexity, as well as the respondents' interest in the task. It might also be the case that some tasks cause distress to respondents (e.g. where there is reference to early death). These aspects of practicality can be assessed by examining the proportion of those approached who agree to participate (i.e. the response rate) and the level of missing data (i.e. completeness).

### 3.2.2 Reliability

Reliability is the ability of a measure to reproduce the same quality adjustment values on two separate administrations when there has been no change in health. This can be over-time, known as re-test reliability, and between methods of administration, known as inter-rater reliability. All measures have some degree of random error. The consequences of more random error is the need for a larger sample size.

### 3.2.3 Validity

The assessment of *validity* is more controversial. The gold standard or criterion test of the validity of a measure intended to reflect preferences would be the extent to which it was able to predict those preferences *revealed* from actual decisions. For welfarists, this would be the preferences of individuals. However, RP methods are not appropriate in the health care field due to the special features of this commodity (Chapter 2). One response to this is to question the value of trying to prove validity at all. This view is reflected in a comment by Williams (1995): *'..searching for 'validity' in this field, at this stage in the history of QOL measurement, is like chasing will o' the wisp, and probably equally unproductive'*. The response of some health economists has been to

focus on establishing the theoretical basis of the measure. This view is typified by the following quote from Gafni and Birch (1995): *"In economics the validity of the instrument stems from the validity of the theory which the instrument is derived from. Thus instead of determining the validity of the instrument itself (the typical case when one uses the classical psychometric approach) one has to establish the validity of the underlying theory."* The theoretical basis of preference-based measures is consumer theory (Johannesson et al., 1996).

The broader extra-welfarist view might be more concerned with social values and this may be different from an aggregation of individual preferences (Loomes and McKenzie, 1989). The criterion test for the validity of a measure as an indicator of social values is less clear than for individual values. The values implied by social decisions have been found to be generate such enormous inconsistencies in the valuation of a comparatively simple outcome of lives saved (Mooney, 1977), that this is unlikely to be a fruitful approach.

Despite these difficulties it is important to examine the validity of a measure. This entails a critical assessment of the two parts of a preference-based measure, the descriptive classification and the preference weights, and empirical validity. The validity of the two components of a measure will indicate the extent to which a measure is able to be a valid cardinal indicator of preferences. The ability to reflect preferences in practice should not be ignored, and therefore ways are proposed for assessing empirical validity.

The remainder of this section sets out the methods used in this review for assessing descriptive validity, the validity of the valuations underlying the preference weights, and empirical validity. The methods proposed incorporate the individualistic and social perspectives and the methodological disagreements in the literature (e.g. on preference elicitation).

Descriptive validity

An accurate description of health is an essential component for a measure to be valid. Published economic evaluations rarely address this issue (Smith and Dobson, 1993). Descriptive validity is assessed in terms of the content, face and construct validity of the descriptions of the health state classifications of the measures.

Content validity

Content validity is defined as the extent to which the items of an instrument are appropriate for the health dimensions being measured (Wilkin et al., 1992). This is important since it determines the content of the utility function i.e. the things which are to be valued. No measure can cover all dimensions and include every conceivable item and there is inevitably a trade-off between completeness and parsimony. When deciding whether to use an instrument, it must be shown to cover or reflect the most important health dimensions and the items should cover the full range of levels of the dimensions and be sufficiently sensitive to significant changes. Claims for content validity typically rest on the comprehensiveness of the instrument and the methods used to generate its dimensions and items.

Face validity

Face validity considers whether the items of each domain are sensible and appropriate. Asking very elderly people, for example, about their ability in vigorous activities (such as running) would be inappropriate. As well as being important for the acceptability of a questionnaire, this determines whether a measure is likely to provide an accurate description of a health state. It is a subjective test, and may be assessed by consulting relevant health professionals, or the patients themselves.

Construct validity

Construct validity is the extent to which a measure correlates with other measures or indicators of health. This test of validity has been developed in psychometrics, but has not been widely used in economics. There are two commonly used approaches.

a) Group comparisons

One approach developed in the psychometrics literature has been to examine whether a measure is able to differentiate between groups thought to differ in terms of their health. The researcher hypothesises an expected pattern of scores by variables such as clinical severity, age, sex, socio-economic characteristics, or recent use of services and then examines the actual distribution of scores (Streiner and Norman, 1989; McDowell and Newell, 1987; Wilkin et al., 1992; Bowling, 1991). Patients with a more severe form of a condition, for example, would be expected to have worse health scores; older people are expected on average to have worse physical health than younger people; and recent visitors to general practice might be expected to have worse scores than those who have not visited recently. This method of group comparisons can never prove the descriptive validity of an instrument since this depends on the hypotheses as well as the measure.

It is important to recognise that these hypotheses may not reflect preferences. Age, for example, may be associated with health, but it cannot be assumed that older people would give a lower valuation for their own health state. Clinical opinion about the severity of a condition or their advice to use health services may be poorly correlated with patients' views (Williams, 1993). A measure may not find a difference between two groups because it is not important in terms of preferences. This might be wrongly interpreted to imply that the health classification is insensitive.

However, the ability of the health state classification to reflect known or expected differences in health can be used to judge the capacity of a classification to describe health. A comparison of groups is therefore best undertaken with the unscored descriptions of an instrument.

b) Convergent validity

This is the extent to which one measure correlates with another measure of the same concept. This has been used to test the ability of the classification of preference-based measures to measure health dimensions in comparison with other widely accepted non preference-based measures such as the Nottingham Health Profile (Whynes and Neilson,

1993). This test also suffers from the problem of circularity in the absence of a 'gold standard'. A strong association between measures may still mean both are invalid. Furthermore, the degree of convergence with a non preference-based measure of health, such as the Nottingham Health Profile, cannot be regarded as a test of ability of an instrument to reflect preferences.

## Valuation

There are three aspects of the scoring system to be addressed: the question of whose values were elicited, the technique for eliciting preferences, and the quality of the data.

### The source of values

Opinions vary in the health economics literature as to whose values should be elicited. This is an important judgement since there is evidence of valuations varying by disease experience, age and education (e.g. Sackett and Torrance,1978; MVH, 1994). It has been argued that respondents who have experienced the health states are in a better position to understand the states (Buckingham, 1993). Another view is that doctors and other health professionals might be thought to have a broader view, and hence be in a better position to understand the relative value of different health states. It has also been argued that it should be a representative sample of the general population for informing the allocation of public resources. There are arguments for all of these constituencies, and they have all been used in past valuation work (Torrance, 1986). The choice of whose preferences or values to use in valuation surveys is an ethical one. It is not for the researcher to decide. It is therefore important for the characteristics of the respondents to be made explicit.

### Valuation technique

The choice of technique for eliciting preferences is examined in Chapter 5. The conclusion is that Visual Analogue Scale (VAS) and Magnitude Estimation (ME) techniques do not generate values which reflect the strength of people's preferences. To elicit preferences it is necessary to confront the respondent with a choice, and therefore either SG or TTO should be used to value health states. There could be a case for using

VAS as a proxy for one of the choice-based techniques by estimating a function for transforming VAS ratings into SG or TTO values. (The theoretical basis for this relationship and the empirical evidence for it are detailed in Chapter 7).

Quality of data

The valuations of the health state descriptions will be based on studies that vary in terms of sample size, and the methods of administering the questionnaires. These have implications for the quality of the data in terms of the reliability of the estimates, and the quality of the respondents' answers. The valuation of the larger generic health classifications needs a method of estimating values for all health states defined by the classification from the valuation of a sample of health states (Dolan et al., 1996).

The reliability of the results of valuation survey should be reported. Large variances in valuations may reflect genuine differences in preferences in the population, but they might be the result of the small sample size in the valuation survey. Where there are significant differences between groups, then it is useful to have specific weights available to conduct a sensitivity analysis.

The quality of the data will depend on respondents' understanding of the task. This should be partly reflected in the logical consistency of their answers. For some health classification systems it is possible to determine a rank ordering. For example, where a health state is better than another state on one dimension but no worse on any other dimension it should be valued at least as highly as the other. The proportion of times respondents' valuations are consistent with this ranking provides an indication of whether respondents understood the task. However, there are no accepted standards of consistency. In some valuation surveys respondents displaying extreme cases of inconsistency are removed (Torrance et al., 1982; MVH, 1994), but this may have implications for the representativeness of the sample.

Some health state classifications are too large to value all health states directly. The Health Utility Index (HUI)-I, for example, has four dimensions and 23 items, while the EQ-5D has five dimensions and 15 items, generating 960 and 243 states respectively,

and the more recent HUI-III with eight dimensions has 972,000 possible states. A sample of health states is valued for such instruments and these are used to estimate values for all their states. This can be done by statistical inference, which involves the use of multi-variate techniques to estimate values for a functional form specifying the relationship between items of the health state classification (e.g. MVH, 1995). The other is an algebraic approach, where individual utility functions are estimated for each dimension, and then aggregated using weights obtained by algebraic solution (Torrance, 1982). These approaches raise important technical issues to be addressed in this review.

Empirical validity

The descriptive content of an instrument and the way it is valued provide the basis for supposing whether or not a measure could generate values which reflect preferences. The ultimate test, of course, is whether the values do so in practice. The difficulties of obtaining revealed preference data in health care were discussed Chapter 2, but there are two less direct tests of empirical validity, one based on stated preferences and the other on hypothetical preferences.

Stated preferences

Given the absence of revealed preference data, an alternative test of the validity of preference-based measures would be a comparison with stated preferences. One application of this approach would be to ask patients to rank health states they have experienced, such as states experienced before and after a surgical operation. A limitation of this, however, is that it would be restricted to testing the ordinal properties of a measure.

Another method is to ask patients to administer one of the preference-based measures alongside a direct method of preference elicitation. The degree of convergence between them would indicate the extent to which preference-based measures generate values that reflect the stated preferences of patients.

For testing the ability of a measure to predict social values, Nord et al., (1993) have argued that: *"..... the validity of the values obtained from different scaling techniques*

*may be tested by asking whether the people from whom the values were elicited actually agree with the consequences in terms of the implied priorities for different health programs."* Nord (1991) has proposed that a version of the equivalence technique, the person trade-off (PTO) technique be used (see Chapter 5 for a review). Assuming that a value of 0.4 has been assigned to state A and 0.8 to state B, then this implies '*the subject is indifferent between making 1 patient in state A well for 2 years and making 2 patients in state B well for 1 year'* (Nord, 1991). Furthermore it has been applied to the Rosser, QWB and HUI by mapping them onto EQ-6D health states. This technique is reviewed in Chapter 5, where it is found to be a difficult technique to apply. It is an interesting approach since it takes a social perspective.

Hypothetical preferences

This is an extension of the psychometric test of construct validity using extreme group comparisons, where the analyst hypothesises expected differences in preferences between groups. It could be hypothesised, for example, that a patient would prefer a less severe condition and hence it should be associated with a higher score. The hypothesis must be chosen with some care, given the reservations already expressed about construct validity. The degree of convergence with a non preference-based measure of health, such as the Nottingham Health Profile, cannot be regarded as a test of the empirical validity of a preference-based measure.

## 3.3 Search strategy and methods of review

This review has been based on a systematic search of the literature undertaken by the Information section at the School of Health and Related Research, the University of Sheffield. The core databases used were MEDLINE, EMbase, Science Citation Index [BIDS] and Social Citation Index [BIDS]. In addition the general economics databases ECONLIT [Silverplatter] and IBIS [British Library Political and Economic Science] were searched. Two approaches were used to identify articles for this section of the review:-

1. by using all permutations of the names of specific scales or instruments presented on Table 3.1; and

2. by performing an author citation search on the original articles that describe the development of each scale or instrument.

A feature of the review was the proliferation of terms for describing the measures which was often compounded by the tendency for a measure or tool to undergo several changes either in its form or simply in the way it is described.

A total of 163 papers were identified by this strategy (Table 3.2). These papers were divided into methodology and applications. The methodological papers (n=92) were those which described the measure and the development of its classification, or presented a review of the measure. Papers reporting the results of administering the measures to patients (n=71) provided the empirical evidence for this review. The empirical papers were summarised, by instrument, in terms of the patient group used in the study, the number of patients, time to complete the questionnaire, response rates, completion rates, and reliability, and whether or not the following were addressed: content and face validity, construct validity, and empirical evidence on relationship to hypothetical or stated preferences (there were no studies reporting revealed preference data).

## 3.4 The review

### 3.4.1 Quality of Well-Being scale

The Quality of Well-Being scale (QWB), formerly the Index of Well-Being, is the oldest of the QALY instruments (though its developers prefer the term "well-year"). The basic structure of the classification and its valuation has remained largely unchanged since the pioneering work of Bush and his colleagues but there have been a number of revisions to its wording, its size and the preference weights (Bush et al., 1982; Patrick et al., 1973a; Kaplan et al., 1976; Kaplan and Anderson, 1988; Kaplan and Anderson, 1990). This review is concerned with the latest published versions of the QWB, although the previous versions are sufficiently related for the earlier empirical work to be relevant to this review.

3.4.1.1 Description

The QWB classification has the three functional scales of mobility, physical function and social function and a list of symptom/problem complexes. The three functional scales have three, three and five levels respectively (Appendix 1.1). The list of symptom/function complexes includes 27 items. The functional states and list of symptom/complexes combine to form 1170 health states.

The patient's functional level on the three scales of mobility, physical function and social function and their symptom/problem complex is obtained from an interview. There are preference weights associated with each function level and these are combined with worst symptom/problem in a simple additive formula to derive the 'index of well-being'(Appendix 1.1).

3.4.1.2 Published literature

There were 32 papers addressing specific methodological aspects of the derivation of the classification, the methods of valuation and the use of the QWB resource allocation decisions and there were 26 published empirical studies using the QWB covering a wide range of conditions. (Table 3.3).

3.4.1.3 Practicality

The questionnaire is administered by trained interviewers. There is a self-completed version, but this method of administration is not recommended since it has been shown to result in the misclassification of health problems (Anderson et al., 1986). It takes between one and two weeks to train interviewers to administer the questionnaire (Read et al., 1987). The interview involves detailed probing of the respondent. The developers claim it can take between 7 and 15 minutes to conduct an interview (Kaplan, 1994), but the range reported in published studies went up to 20 minutes (Bombardier and Ramboud, 1991).

Few studies have formally reported response rates. In one study with older adults, the response rate was 68.2%, but 100% was achieved in the study of COPD patients (Kaplan et al., 1989). The rate of completion was 93% and 100% in each of these

studies. Andresen et al. (1995) found it was more complex than the Sickness Impact Profile and the SF-36, and Wu et al. (1990) and Bombardier and Ramboud (1991) thought it a complex instrument.

### 3.4.1.4 Reliability

The only published article reporting on re-test reliability was an assessment of the inter-day reliability (Anderson et al., 1989). The authors used the results of five empirical studies which found that assessments one day apart had correlations of 0.78 to 0.99 and the majority were in excess on 0.9. However, the ability of this study to assess re-test reliability must be questioned because the data were obtained retrospectively in one block rather than prospectively.

The reliability of the interview method has been examined by testing the accuracy of assignment against a recording of the interview. Ninety six percent were found to be classified correctly. There were no papers on inter-rater reliability. A comparison of self versus interviewer modes of administration found correlations of 0.98, but the authors believed this masked some important differences owing to false self-reporting associated with the self-completion (Anderson et al., 1986).

### 3.4.1.5 Descriptive validity

<u>Content and face validity</u>

The first version of the classification was based on items from a review of the literature and of survey instruments used over the previous decade (including the U.S Social Security Administration Survey of the disabled and the Health Interview Survey). The developers claimed the function scales and symptom and problem item list were exhaustive. The specific reasons for the choice of mobility, physical function, social function and the symptom/ problem list have not been published. Some of the function levels and the items in the list of symptoms were merged and others were excluded in subsequent versions of the instrument (Kaplan,1989). These changes were based on experience from using the instrument or the results of the valuation. Items in the symptom/problem list found to have approximately the same rating by respondents were

combined[1] and four items were added to the list of problems and symptoms. Other items can be added to the list.

The QWB seems to be comprehensive in its coverage of function and symptoms or problems, but it has been observed that it is less comprehensive in mental health (Read et al., 1987). Mental health is not assessed as a separate dimension in the QWB, though the most recent version has a symptom/problem called 'excessive worry or anxiety'. The developers believe mental health affects function in the same way as physical health, and should not require its own dimension. This ignores a substantial body of work which shows mental health domains, such as depression and anxiety, to be distinct constructs (Ware et al., 1984). The QWB also excludes those aspects of health concerned with social support and friends. The social function dimension is limited to participation in work and attendance at school and not leisure activities.

Researchers have expressed concern at the insensitivity of the classification (Tandon et al., 1989; Liang et al., 1990). In the latest version, two of the three functioning scales have only two dysfunctional levels and this would seem to permit little scope for measuring change. Kaplan et al. (1976) have argued that it is the symptom/problem complex list which makes the instrument sensitive. Furthermore, given the multi-collinearity between the components of the QWB, it is not appropriate to separate out the sub-scales. The list of CPX items is indeed very extensive, but at face value the items do not seem very sensitive since they are dichotomous. There is no allowance for the intensity or frequency of the symptom or problem. For example, you either have, or do not have, trouble with sleeping, and such a dichotomy seems unlikely to measure small but potentially important improvements in sleeping. This may be less improtant in practice because the scoring of this domain works by selecting the worst symptoms or problems associated with a given state of ill-health, and thereby achieve a finer gradation in practice. For example, the worst problem may switch from troubled sleep to pain in the ear following a successful intervention. The ability of this scoring algorithm to overcome the insensitivity of the descriptors is an empirical issue.

---

[1] The version in Kaplan and Anderson (1988) combines items 3, 4, 5 & 6 from Kaplan et al. (1976) into a single item.

There has also been re-wording of the items from the original version, mainly to replace items about capacity with those concerned about behaviour and actual performance (Kaplan et al., 1976; Kaplan and Anderson, 1990). This contrasts with the HUI classification which is concerned with capacity. Kaplan et al. (1976) have argued that asking about behaviour and actual performance avoids the respondent having to make difficult judgements about what he/she could do.

The wording of the items in QWB seems straightforward and in most cases reasonably clear. Some of them are lengthy, however, and combine quite disparate things. The social activity scale, for example, combines work with self-care activities. In the symptom/problem list, one item combines "*hands, feet, arms or legs either missing, deformed, or paralysed*". Another combines "*pain in ear, tooth, jaw, throat, lips and tongue*", *with* "*runny nose*". These were combined on the grounds that they have been equally valued, but it is questionable whether they make much sense together.

Construct validity

Thirteen of the 23 studies listed in Appendix 2.1 were found to report results on the construct validity of the QWB. The QWB has been found to be significantly correlated with the general health status measures of the SIP (Hornberger et al., 1992; Read et al., 1987) and the SF-36 (Andresen et al., 1995) and with the condition-specific Arthritis Impact Scale (Kaplan et al., 1984), the Functional Status Index (Ganiats et al. 1992) and the Karnovsky Performance Scale (Wu et al., 1990). Kaplan et al. (1995) and Orenstein et al. (1989, 1990) have also claimed to have demonstrated convergent validity in terms of correlation with various clinical measures used in COPD and cystic fibrosis, including respiratory function (e.g. $FEV_1$ ), and exercise tolerance. These studies have provided consistent evidence of the convergence of the QWB score with measures of function. The doubts raised earlier about its coverage of mental health, however, found some support from the study by Andresen et al. (1995) who found it to be poorly correlated with emotional and psychological measures of health in a comparison of measures in healthy older adults (i.e. the SIP, SF-36, and positive affect scale), though

Kaplan et al. (1995) found it was significantly correlated with the Becks Depression Inventory.

Holbrook et al. (1994) found the overall QWB score significantly improved in trauma cases between discharge and a three month follow-up. The authors also noted that the QWB continued to identify limitations in this patient group, whereas the more condition-specific Functional Status Index did not, and they therefore concluded that the QWB was a more sensitive measure of function. The QWB was also found in this study to be as sensitive as other measures of function i.e. the Hospital and Anxiety Questionnaire and The Keitel Assessment (Bombardier et al., 1986). In contrast, Laing et al. (1990) found that the functional scales of the QWB were not able to detect change in orthopaedic patients following surgery, in comparison with four other health status instruments, though the overall index did detect a change. The QWB also failed to detect a difference between congestive heart failure patients receiving standard therapy and those allocated to placebo, which had been shown by a set of patient-completed symptom scales and the physician assessed Spitzer Quality of Life scale (Tandon et al. 1989). The individual components of the QWB were unable to find a difference between these groups. There was further evidence of the insensitivity of the QWB to psychological outcomes in a study by Calfas et al. (1992) who evaluated the effects of a cognitive-behavioural intervention in osteoarthritis patients compared to a control group. Differences were found in the Beck Depression Inventory at one year, but these were not reflected in the QWB.

### 3.4.1.6 Valuation

A stratified random sample of 343 health states was selected and divided into eight booklets. These booklets were each valued by approximately 100 respondents[2] using a version of the visual analogue scale. Respondents were asked to place each state into one of 15 numbered slots defined by a scale from zero to 16 where zero was death and

[2] These figures were taken from Kaplan and Anderson (1988). It is unclear from published sources whether these 343 health states are from the revised classification or the longer version in use at the time (e.g. the original survey included age in the health state descriptions).

16 optimum health[3]. The results were transformed onto a zero (death) to one (optimal health) scale. Linear statistical models were fitted to the transformed mean and median health state values to estimate weights for the levels of each function and the list of symptoms and problems.

The 866 respondents were selected to be representative of the general population of San Diego. The developers argued that the results are generalisable since they found background variables to make little difference to the mean valuations (Kaplan et al., 1976) . Balaban and colleagues (1986) found the weights from a sample of rheumatoid arthritis patients to be very similar. However, these samples would not have included the full range of background variables that would be found over a wider and more diverse population, such as in the UK. There is little reported on the quality of the data from these surveys.

The use of VAS to value health states can be criticised for not being a choice-based technique. Kaplan and his colleagues have argued strongly in favour of VAS over other techniques as a measure of preferences, but these arguments have been drawn principally from the psychometric literature (Kaplan and Ernst, 1983). There is no basis in economic theory for the claim that VAS can reflect preferences (see Chapter 3)[4]. Nord (1993) argues that the QWB weights imply *'too low equivalent numbers for trivial treatments compared to treatments for severe conditions'* and this has been shown to lead to some absurd policy implications in the Oregon experiment with setting priorities according to cost per well year (Nord,1993).

It is difficult to judge the validity of the statistical model used to derive the preference weights. The authors have reported an overall $R^2$ in excess of 0.96, but they failed to provide detail about the standard errors associated with the coefficients, the results of any diagnostic tests (such as homogeneity and normality in the error term) or the results of other model specifications (including possible interactions). There have been two models reported on the San Diego data, but no evidence given for the superiority of the

---

[3] As described by Patrick et al. (1973) in an earlier publication.

[4] It is interesting to note that the results from the San Diego survey seem to exhibit the common tendency in VAS data to cluster near the middle, since there was little variability between states.

more recent model (Kaplan et al., 1976 and Kaplan and Anderson, 1988). Anderson (1982) has shown that the earlier model implied some counter-intuitive rankings of the levels within scales. A movement from "moved own wheel chair without help'' to "walked with physical limitations'' actually resulted in a reduction in the overall score. This could be due to mis-specification in the model, such as the existence of interactions. More formal testing of the model is required than is currently available.

### 3.4.1.7 Empirical validity

Out of the studies listed in Table 3.3, five were found to report evidence relevant to assessing the empirical validity of this instrument. Four of these studies reported evidence of agreement between QWB scores and hypothetical preferences. The richest data set has been generated from a study by Fryback and colleagues who administered the QWB alongside a questionnaire recording the number and type of medical conditions. As expected, QWB scores were found to decline as the number of medical conditions increased. This confirmed results published by developers of the QWB (Kaplan et al., 1976), who found a correlation of -0.36 between the number of conditions and QWB score at the individual level. Furthermore, age-specific scores were found to be consistently lower in adults with arthritis, severe back pain, or sleeping disorder compared to those without these conditions. For adults with the less severe condition of hypertension the differences were smaller or zero. Kaplan and his colleagues also found the score to be correlated with the number of recent physician visits. The finding by Holbrook et al. (1994) of QWB scores improving in patients recovering from trauma were also in line with expectations. Finally, a study by Kaplan et al. (1995) found QWB scores were significantly different between HIV severity groups.

Validity against stated preferences has been reported in the form of convergence with directly administered TTO and SG questions. In the survey by Fryback et al. (1993) TTO and the QWB score were found to correlate by 0.41 and in a comparison by Hornberger and colleagues the correlations were 0.31 and 0.42 for TTO and SG respectively.

3.4.1.8 Overview - key points

- Interview administration makes this the most time-consuming and expensive of the preference-based instruments (though substantially less than many routine medical tests).
- No assessment of re-test or inter-rater-reliability has been found.
- The descriptive system seems comprehensive in relation to the function and symptoms, but there is little on mental health problems.
- Evidence of descriptive validity has been primarily of correlations between the QWB score and measures of health status. There is some evidence of the insensitivity of the function scales.
- There is no theoretical support for the method of valuation, namely VAS. The model used to estimate the published weights has not been subject to rigorous econometric testing.
- Scores have been in line with prior expectations of preferences and have correlated significantly with direct preference measures.

### 3.4.2 Rosser Classification of illness states

The classification was developed by Rosser and others in the 1970s as a generic measure of hospital output (Rosser and Watts, 1972; Rosser and Kind, 1978). The content of the classification has remained largely unaltered, though different methods of administration have been developed, including a self-complete version. In the 1980s, it became the most widely used instrument for deriving QALYs in the UK.

3.4.2.1 Description

The Rosser classification has two dimensions, Disability and Distress, with eight and four levels respectively (Appendix 1.2). Together these categories define a total of 29 health states (being unconscious suggests no distress and therefore three have been excluded). In early applications clinicians classified patients using a brief one page reminder of the meaning of disability and distress (which includes pain and mental disturbance). Other published studies have asked clinicians to place their 'average'

patient on the classification before and after treatment (e.g. Williams, 1985). More recently a self-completed instrument called the Health Measurement Questionnaire (HMQ) has been developed for classifying patients onto the classification (Gudex and Kind, 1988). Researchers have also mapped patients onto the classification from other health status questionnaires (e.g. Gudex and Kind, 1988). The original values for the classification are presented in Appendix 1.2.

### 3.4.2.2 Published literature

There were 21 papers on the development of the classification and its valuation, reviews, and discussions of its application to NHS decision-making. Twenty three papers reported its application on patients, though two were reporting results from the same study (Table 3.2 and Table 3.4).

### 3.4.2.3 Practicality

Clinical assessment takes just 10 seconds and can be done as part of routine practice (Rosser, 1988). The most common method of administration has been the Health Measurement Questionnaire (HMQ), by either patient self-completion or interview. The self-completed HMQ offers a comparatively easy method and its developers claim it takes no more than 10 minutes to complete. By interview administration it takes somewhat longer, and in the one study reporting timings it took 30 minutes (Magee et al., 1992). Response rates in patient groups ranged between 76-95%. Completion rates were 87% and 95.5% in the two studies reporting them, but in a number of other studies the completion was 100% by implication.

### 3.4.2.4 Reliability

In the initial work with the classification, inter-clinician agreement was high (Rosser and Watt, 1972). This result was repeated with ward nurses (Benson, 1978). In a more recent study by Bryan and colleagues (1991) on chiropody patients, however, substantial disagreement was found between clinicians. Significant differences have been found between clinician and patient-completed HMQs (Petrou et al., 1992;

Whynes and Neilson, 1993). More evidence is required on the re-test reliability of results generated by the HMQ.

Questions have been raised about the assignment of patients onto the Rosser classification by mapping from other questionnaires. Drewett et al. (1992) believed this explained the large variation between the valuation of the health gain from knee replacements from their studies and those published elsewhere (Williams, 1985). Coast (1992), however, found reasonable agreement between the 13 raters who undertook a transformation from one questionnaire to another, though she had considerable doubts about the validity of the exercise.

### 3.4.2.5 Descriptive validity

Content and face validity

Two dimensions limit the comprehensiveness of the Rosser classification, though the dimensions describe more than one domain of health. Disability is intended to assess observable factors, such as the patient's mobility and self care, and Distress assesses subjective aspects such as pain and distress. Energy, mental health and many other symptoms of disease are not included in their own right, though it might be argued that they will be reflected in one or both of the dimensions. The reasons for choosing the two dimensions are not reported.

The descriptions were developed from asking 60 doctors to identify those features they took into account in assessing illness severity (Rosser, 1988). The dimensions have been criticised for being difficult to interpret (Elvik, 1995). Pain and mental disturbance are both encompassed by the Distress dimension (Gudex and Kind, 1988) and yet these are very different aspects of health. There is also ambiguity in the wording of the levels of the disability dimension. It is not clear, for example, that level four is unambiguously better than five. Gudex and colleagues (1993) suggest difficulties may arise, for example, from the large amount of text in level 5 of disability. The notion of social disability is also ambiguous and this is reflected in the substantial inconsistencies found between median health state values and the logical ordering of health states (Gudex et al., 1993).

At face value, the categories of each scale of the Rosser would seem very crude. The instrument was originally developed as a measure of hospital output and hence intended to measure large changes. The developer of the instrument has since argued that it is not suitable for trials (Rosser, 1988) and hence it will be too blunt to assess strength of preference for the more subtle differences arising between hospital treatments, and for most treatments provided in primary and community settings.

The face validity of the method of transforming responses on the HMQ onto the Rosser classification has also been questioned by Bryan and colleagues (1991) and Carr-Hill and Morris (1991). According to the assignment rules, a person in category IV has difficulties with washing, dressing, eating and drinking and using the toilet, and his/her social life, seeing friends or relatives, hobbies/leisure activities and sex life are all affected by health, and yet this person is assumed to be able to do all his or her usual activities. The mapping of patients onto the classification from other questionnaires has been found to be of questionable value since the process is based on a large number of arbitrary assumptions (Coast 1992; Drewett et al., 1992).

Construct validity

Studies have found the classification to be sensitive to the outcomes of hip and knee replacement (Petrou et al., 1992; Drewett et al., 1992; Chan and Villar, 1996), cardiac surgery (Kallis et al., 1993), elective surgery for abdominal aortic aneurysm and chiropody services (Bryan et al., 1991). The overall index was also able to distinguish between end stage renal patients on transplant and dialysis (Gudex 1995). These results contrast with the study by Donaldson et al. (1988) who found the Rosser was unable to detect changes in a trial of long-term care for elderly people, when a majority of patients had changed according to measures of disability and psychological well-being regarded as more suitable for this group (Crichton Royal Behavioural Rating Scale and the Life Satisfaction Index respectively). A study of patients with knee problems found the index was unable to show differences between the patient group and the general population, which had been found by both SF-36 and EQ-5D (Hollingworth et al., 1995). Furthermore, it was unable to show the improvements at six months indicated by these other instruments. Hollingworth et al. have argued that this may have been due to the

small range of values in the original valuation matrix, rather than necessarily a fault of the classification.

The Rosser was found to correlate with the Nottingham Health Profile (NHP) dimensions (Whynes and Neilsen, 1993; Kind and Gudex, 1994), the GHQ-12 (a measure of psychiatric disturbance; Kind and Gudex, 1994) and the Dallas pain questionnaire (Launois et al., 1994). The Disability scale was found to correlate most strongly with the mobility scale of the NHP, then pain and energy. For Distress, the strength of correlation was strongest for Emotional Reaction. However, it would seem that the pain scale of the NHP was more strongly associated with Disability than Distress. This highlights the ambiguity of the concepts underlying the Distress dimension.

3.4.2.6 Valuation

Published work using Rosser has been limited to the original valuation study undertaken by Rosser and colleagues. Seventy respondents were asked to rank six 'marker states' (chosen to cover the full range of the classification), and then value five of them in terms of the 'least ill state' using a version of magnitude estimation. The remaining 23 states were ranked and valued in the same way, as well as death. Respondents were asked to consider the implications of their answers in terms of the allocation of resources between patients in the different health states. Responses were found to be reliable at re-test and between observers (Rosser and Kind, 1978). The results were averaged across all 70 respondents and transformed onto a scale from zero to 1.0, where zero was set at death and one full health. Separate matrices of values have been produced for each of the professional and patient groups.

There has been concern at the unrepresentativeness of the 70 respondents and the small numbers. These could be important, given the finding that valuations varied between groups (Rosser and Kind, 1978). Magnitude estimation has no theoretical basis in economics, and cannot be regarded as appropriate for economic evaluation (Johannesson et al., 1996). However, the discussion of the resource use implications of their valuations during the interview provided a framework of choice, and Nord (1992)

has argued that the values in this matrix of values appeared to be more consistent with his equivalent numbers test than those from other instruments.

The revaluation of the Rosser classification by TTO could have provided a theoretically more acceptable method to use in economic evaluation (Gudex et al.,1993). The matrix of values differs considerably from the original. The values were lower and were found to have important implications for the cost-effectiveness of interventions in terms of their cost per QALY ratios. There were some important 'reversals' in the ordering of some states and particular problems arose with the valuation of states worse than death. The developers did not believe these TTO valuations to be better than either of the new VAS and ME valuations. They have recommended that those wishing to conduct QALY analysis using the Rosser choose between the original ME matrix, a new ME matrix, or a matrix based on a 'synthesis'of VAS, ME and TTO. There is no theoretical basis for believing that the values from either of the ME matrices or the synthesised matrix reflect preferences on a cardinal scale.

3.4.2.7 Empirical validity

The studies showing the ability of the Rosser to detect the expected improvements following hip and knee replacement (Petrou et al., 1992; Drewett et al., 1992; Chan and Villar, 1996), cardiac surgery (Kallis et al., 1993), elective surgery for abdominal aortic aneurysm, and chiropody services (Bryan et al., 1991) all provide evidence of the ability of the index to reflect hypothetical preferences. The higher index score of transplant patients compared to those on dialysis also confirmed earlier research findings that patients prefer transplants (Sackett and Torrance, 1978). The study by Hollingsworth et al. (1995) of patients with knee problems found the index was unable to show differences between the patient group and the general population, or improvements at six months found by the EQ-5D.

Nord and his colleagues (1993) compared the values of the original Rosser matrix to the responses to PTO questions. Along with the QWB and HUI-I, it was mapped onto two EQ-6D health states. The Rosser generated values nearer to the PTO valuations than the other preference-based measures, and therefore Nord and colleagues argued that it better

reflected social preferences. This study had a number of methodological weaknesses in terms of reliance on dubious mapping procedures, and small samples. Furthermore, the PTO values resulted in an illogical ordering of the two EQ-6D health states.

3.4.2.8 Overview - key points

- Both clinical assessment and the patient completed HMQ are practical methods of collecting descriptive data.
- There is little evidence on reliability of these methods.
- Two dimensions provide only limited coverage. The descriptions partly overcome this by tapping more than one domain, but this results in ambiguities in the ranking of the levels of disability.
- There is evidence suggests that the Rosser classification is sensitive to large changes, such as those associated with major surgery in hospital, but it is not designed for measuring more subtle changes. There is evidence of insensitivity in the classification.
- There is no justification in economic theory for the original method of valuation as a measure of preferences, nor the recommended 'synthesis' of these values and the new ME and TTO values.
- Evidence on hypothetical preferences in group comparisons, but insensitivity found due to the original scoring algorithm.

### 3.4.3 Health Utilities Index

The Health Utilities Index (HUI) was devised by Torrance and colleagues (1982). The earliest version, now known as HUI-I, has been succeeded though not replaced by two revised classifications, HUI-II and III (Torrance et al., 1995; Feeny et al., 1995). HUI-III is closely related to HUI-II but both differ substantially from HUI-I. All three versions are reviewed here.

3.4.3.1 Description

HUI-I is composed of four attributes or dimensions (physical function, role function, social function and health problems), with four to eight levels each, defining 960 unique

health states (see Appendix 1.3). HUI-II has seven dimensions: sensation, mobility, emotion, cognition, self-care, pain and fertility, with three to five dimensions and defines 24,000 states in all. HUI-III is an adaptation of HUI-II. The number of dimensions has been increased to eight and includes vision and hearing as separate dimensions, along with speech, ambulation, dexterity, emotion, cognition and pain. Fertility was removed. The number of levels has been increased to between five and six, and it defines 972,000 health states. Patients are assigned to the classifications from a self-completed questionnaire, or from interview face-to-face or by telephone. Patients have also been mapped onto the levels of each dimension of the HUI-I from responses to other health measures (Gold et al.,1996).

A total TTO utility value is obtained from HUI-I by inputting the dimension level weights shown in Appendix 1.3 into the following formulae:

$$U = 1.42X (P_i \times R_i \times S_i \times H_i) - 1.42$$

where $P_i$ equals the preference weights for the level on physical function, $R_i$ is the preference weight on role and so forth. A similar multiplicative algorithm has been estimated for HUI-II. The weights for HUI-III have not been published at the time of writing this thesis.

3.4.3.2 Published literature

Out of a total of 21 papers identified in the search, 11 were methodology; presenting descriptions of the HUI and its origins, reporting the results of the valuation surveys, and describing the application of multi-attribute theory to the classifications to derive the algorithms for valuing all health states. Two papers were concerned with HUI-I, four with HUI-II, and five with HUI-II and III. There were ten empirical studies using one of the HUI classifications (Table 3.5). HUI-II has been the most widely used to date, with seven papers. Eight of the 10 applications of the classifications have been with young survivors of low birthweight or various forms of cancer, reflecting the origin of the instruments. The remaining three have been adult populations. Only two of the 21 publications have come from research groups outside McMaster University.

3.4.3.3 Practicality

HUI-I was originally administered by home interview (Boyle et al., 1983). HUI-II has been administered prospectively by health professionals who knew the patient; by interview with patients and/or their parents face-to-face and by telephone; and by a self-completed version mailed to respondents. Patients have also been assigned retrospectively using other health assessment data (Saigal et al., 1994). The developers now recommend a 15 item questionnaire for self-completion or interview administration.

Two studies report that administration took 1-2 minutes by health professionals known to the patient and 5 minutes for interviews of patients and their parents (Billson and Walker, 1994; Barr et al., 1993). Response and completion rates are rarely reported. Some studies seem to imply 100% (e.g. Barr et al., 1993). Reported response rates vary between 79-100% and completion between 96-100%. The figure of 79% was achieved in a routine clinic where there were a number of reasons for the low rate that were unrelated to the willingness on the part of the patient (Billson and Walker, 1994).

3.4.3.4 Reliability

In terms of inter-rater reliability, discrepancies were found in the assignment of patients onto HUI-II, though these usually involved one dimension level (e.g. 39% disagreement was found by Feeny et al., 1993 and 30% Barr et al., 1994). There did not appear to be any systematic pattern to differences between professionals, but they were found to identify fewer problems than the patients or their parents. Barr et al. (1994) argued that this discrepancy arose because patients and parents were better informed than the health professional, particularly in the subjective areas such as pain and emotions. The developers recommend that a common method of assessment is used throughout a study.

There has only been one study of re-test reliability and this was in a general population survey using the HUI-III (Boyle et al., 1995). Individual responses were found to be stable between tests for six dimensions, the exceptions were speech and dexterity (Boyle

et al., 1995). The instability of these two dimensions was claimed to be due to their infrequent reporting in the populations surveyed. It is not clear why infrequency should result in instability. The re-test reliability (12-49 days apart) of a provisional overall HUI-III index score was found to be 0.77 (intraclass correlation coefficient).

### 3.4.3.5 Descriptive validity

Content and face validity

HUI-I was devised by Torrance and colleagues (1982) to assess the outcome of survivors of neonatal intensive care. It was designed to include the range of health problems likely to be experienced by long term survivors of neonatal intensive care and based on the multi-attribute framework developed by Bush and colleagues (Feeny et al., 1995). It covers the physical, mental and social domains mentioned in the World Health Organisation's definition of health (see Chapter 2). Energy and pain are included as levels within the 'health problem' dimension.

The descriptions are quite lengthy, and often combine more than one domain of health. The combinations are logical for physical function, but less obvious for the role dimension which combines self-care and role activity, where ability to eat, dress and bathe are combined with limitations in playing, going to school and so forth. The aggregation of emotional well-being with social activity also does not seem appropriate and creates further ambiguities in the ranking of levels. The health problems dimension is a mix of problems, and with no obvious ordinality.

HUI-II was initially designed to assess health status in long-term survivors of childhood cancer. It was based on a review of the literature which identified 15 potential attributes. These were presented to parents and children who were asked to identify the six which were most important to them (Cadman et al., 1984). The number of levels was also based on a review of existing instruments.

The health state classification of HUI-II is very different from HUI-I. It includes cognition and fertility, pain is made into a separate dimension, the health problems

dimension has been dropped, and it entirely excludes work and social function. The authors argue this is a generic measure of health. However, its content reflects the patient group for whom it was originally designed. The wording of the content of the instrument quite explicitly aimed at children (e.g. "ability to see, hear and speak normally for age"; "learns and remembers school work normally for age"). The inclusion of fertility indicates a more condition-specific and it does not appear in any other generic measure of health.

The authors argue for a 'within skin' definition, which is only concerned with impairment and disability and not handicap. Social and role activities are a consequence of people's preferences and overall choice set, and hence should be excluded from a pure description of health. However, the classification in HUI-II is not entirely 'within skin' since some dimensions (mobility, self-care, sensation and cognition) contain references to independence from help and mechanical aid, which are likely to be influenced by a person's setting.

The dimensions of HUI-II are focused on single attributes and hence are less likely to generate the ambiguous rankings of the previous version. Furthermore, the statements are shorter. The exception to this is emotion, where the items include a listing of moods e.g. *''often fretful, angry, irritable, anxious, depressed, or suffering night terrors''*. These are a very mixed set of emotions. One research team found it necessary to simplify this dimension further in order to administer the questionnaire (Kanabar et al., 1995). The descriptions also reinforce the impression that this instrument is intended for children.

Experience with HUI-II resulted in the developers making a number of revisions, and to enhance its relevance for an adult population. The replacement of self-care by dexterity has improved its independence from other dimensions, though this has resulted in the removal of key functions such as bathing, dressing and eating. The disjoining of vision, hearing and speech into separate dimensions makes the HUI more comprehensive and a much larger classification. However, the mental health dimension can be criticised for

having simple statements relating to degrees of happiness, rather than mental problems such as depression or anxiety.

The influence of the earlier work on survivors of childhood cancer and neonatal intensive care is evident in HUI-III. The dimensions are those which are important to parents in regard of their children, such as speech and cognition, but there is rather less emphasis on mental health and nothing on energy or sleep, which are likely to be of more relevance to older people.

Construct validity

Most of the published evidence to date comes from applications of HUI-II to survivors of childhood cancer. Among fifty patients who had acute lymphoblastic leukaemia (ALL) in their childhood, Barr et al. (1993) found a greater burden of ill health amongst patients who had higher risk conditions (70% had a problem compared to 40% in the lower risk group) and as would be expected, this difference was most noticeable on the emotion and cognitive dimensions. In a study of only 10 brain tumour patients, differences were found compared to a normal population in terms of cognition (Barr et al., 1994). Differences have also been found in 156 patients who had a childhood brain tumour between those being treated and those no longer on treatment (Feeny et al., 1993). The HUI-II has also been shown to be able to discriminate between extremely low birth-weight children and a random sample of children (Saigal et al., 1994). There have been concerns about its sensitivity since in these patient groups a large proportion were found to have no problems (Barr et al., 1994), and in another comparison of ALL patients with the general population it was not possible to find differences (Feeny et al., 1993b). There have been no published studies of the construct validity of HUI-III.

Given the limited range of conditions on which it has been tested, the developers acknowledged in a review in 1995 that it is not possible to establish the sensitivity of the HUI classification and that "to date, there is only fragmentary evidence of the ability of the HUI-II or III system to capture change in health status" (Feeny et al., 1995).

3.4.3.6 Valuation

HUI-I and II were valued by random samples of 87 and 203 parents of school children (Torrance et al., 1992). Torrance and his co-workers used a well-tested set of visual aids for eliciting values, and achieved good levels of reliability in the surveys (Torrance et al., 1982). The response rates in the surveys to value HUI-I and II were 75% and 72% respectively, though a large number of respondents were excluded because of missing data, poor quality interview or evidence of confusion with the valuation tasks. These problems resulted in the exclusion of a further 22% and 29%. HUI-III has been valued by a sample of 504 adults from Hamilton, Ontario.

The HUI-I and HUI-II were valued by random samples of parents of school children from Hamilton, Ontario, since this was the constituency of interest in these studies. The generalisability of valuations based on comparatively small samples of parents to other populations has not been established though valuation work with an earlier version of the HUI-II version on a sample of the general population found the valuations to be similar to those from a sample of parents, but the samples contained only 32 in each group (Cadman et al., 1984). HUI-III has been valued using a stratified random sample of 504 individuals in Hamilton.

The initial choice of TTO was based on the premise that it was a good proxy for SG. Torrance and colleagues now acknowledge this is not the case and for HUI-II and III have used SG (Torrance et al., 1995). The HUI I and II were valued using a transformation of VAS ratings to TTO or SG using estimated power functions. The difference between VAS ratings and SG utilities is assumed to be a person's attitude to risk. The validity of this transformation has been questioned in the literature (see Chapter 5). Other researchers have shown a linear model to provide as good a fit as a power specification (Loomes, 1993) and indeed, in a recent study using data from the MVH study found the quadratic and cubic linear models to perform better than Torrance's power function (Dolan and Sutton, 1997). Results from similar tests have not been published on the HUI data, although there is evidence of problems with the model from the substantial divergence between actual SG values for HUI-II states and the predictions from the transformation of the predicted VAS values (i.e. -0.06 to 0.34

across 4 states; see Torrance et al. (1992)). Finally, there are major theoretical doubts about whether attitude to risk is the only difference between VAS and SG. As reported in Chapter 5, there are doubts as to whether VAS can be regarded as anything more than a measure of ordinal preferences.

An important feature of the HUI has been the application of Multi-attribute Theory (MAUT) to derive its weights. MAUT substantially reduces the valuation task by making simplifying assumptions about the relationship between dimensions. The first task was to value the levels of each attribute, to derive a set of single attribute utility functions. A sample of multi-attribute states is then valued and an overall function is calculated by solving a system of simultaneous functions. This is made possible by assuming, for example, an additive functional form where the dimensions are assumed to be independent. This permits no interaction. This was found to be invalid, and the multiplicative function[5] has been used to value the HUIs. The multiplicative function

---

[5] **Types of multi-attribute utility theory models**

**Additive:**

$$u(X) = \sum_{j=1}^{n} k_j\, u_j(x_j) \qquad (1)$$

$$\text{where : } u(X) = \sum_{j=1}^{n} k_j = 1 \qquad (2)$$

**Multiplicative (see note):**

$$u(x) = (1/k)\left[\prod_{j=1}^{n} (1+k\, k_j\, u_j(x_j)) - 1\right] \qquad (3)$$

$$\text{where } (1+k) = \prod_{j=1}^{n} (1+k\, k_j) \qquad (4)$$

**Multilinear:**

$$\begin{aligned} u(x) = {} & k_1\, u_1(x_1) + k_2\, u_2(x_2) + \ldots \\ & + k_{12}\, u_1(x_1)\, u_2(x_2) + k_{13}\, u_1(x_1)\, u_3(x_3) + \ldots \\ & + k_{123}\, u_1(x_1)\, u_2(x_2)\, u_3(x_3) + \ldots \\ & + \ldots \end{aligned} \qquad (5)$$

where the sum of all k's equals 1.

Notation: uj(xj) is the signal attribute utility function for attribute j.

permits a very limited form of interaction between dimensions which assume the inter-dependency to be the same between all dimensions and for all levels of each dimension. For HUI-III the plan is to estimate the less restrictive multilinear functional form[5].

The application of MAUT enables the assumptions of the different models forms to be tested. However, it is not based on the ability to predict values, and does not provide a method of systematically testing the errors in its predictions. The predictive validity of HUI-II has so far only been examined for four health states and large difference were observed. This is too few observations to be a sufficient test of its predictive validity. There has been a comparison of the MAUT approach with a statistical one in a study of job choice by Currim and Sarin (1984). They found the statistical approach substantially outperformed the algebraic: the correlation between actual and predicted choices over jobs (with different mixes of attributes) was 0.16 for the algebraic method and 0.64 by statistical inference from SG utility values. More evidence is required on the ability of this method to predict health state values.

3.4.3.7 Empirical validity

The HUI has not been widely used, and there is very little evidence on empirical validity (Torrance et al., 1995). There are only two published studies and both of these relate to HUI-I. The original application of HUI-I to survivors of neo-natal intensive care found those who had a lower birthweight had a lower quality of life (Boyle et al. 1983). The other study has mapped responses from a general population health survey onto the HUI-I and generated a score for 10,163 persons (Gold et al. 1996). These scores were associated with a number of medical conditions, education, income and ethnicity. This provides some evidence in terms of hypothetical preferences.

3.4.3.8 Overview - key points

---

u(x) is the utility for health state x, represented by an n-element vector.
k and kj are the model parameters

Note: The multiplicative model contains the additive model as a special case. In fitting the multiplicative model, if the measured $k_j$ sum to 1, then k = 0 and the additive model holds.

*Source: Torrance et al. (1995)*

- The 15 item questionnaire is brief and easy to use.
- There is no evidence on re-test reliability in patient groups. The same method of administration must be used to undertake comparisons.
- The content of HUI-II and III would seem to be better than HUI-I.
- The HUI-II and III are comprehensive on physical health, but weaker in terms of mental health, and exclude 'social' health. The content of the HUI-II and to a less extent HUI-III, reflect concerns with the health of children.
- Applications have been very limited to date (mainly HUI-II on survivors of childhood cancer). There is some suggestion of possible insensitivity in HUI-II.
- The validity of the methods of valuation depends on a transformation of VAS to SG and the unproven predictive properties of MAUT.
- There was very little evidence (for or against) empirical validity.

### 3.4.4 The 15D

This measure originally had 12 dimensional classification, but it has been revised to 15 dimensions (Sintonen and Pekurinen, 1993). Further revisions have been made to the dimensions to form the 15D.2 and this is the recommended version for future applications (Sintonen, 1994a & b). Evidence from both versions of 15D is reported here, since the 15D.1 is sufficiently similar to its successor to be relevant[6].

#### 3.4.4.1 Description

The dimensions of 15D are mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort and symptoms, depression, distress, vitality, and sexual activity. Each dimension has 5 levels and hence the classification is able to define many billions of health states (Appendix 1.4). Patients are classified by a self-completed questionnaire where respondents are simply asked to indicate their level of health on each of the 15 dimensions.

---

[6] An instrument has been developed for measuring health-related quality of life in adolescence based on the 15D, but this review has been limited to measures of adult health(Apajasalo et al., 1996).

Health state values are estimated from a simple additive formula, where a value is assigned to each dimension level, and these are multiplied by a weight representing the relative importance of that dimension and summed to derive a single index. The scoring algorithms for 15D.1 are presented in Appendix 1.4. The final scoring and weighting algorithms for the 15D.2 are available from Professor Sintonen.

3.4.4.2 Published literature

The search identified just nine publications, including six refereed articles, a book chapter and two working papers (Table 3.2). Five of these publications were concerned with methodology, one with the 12D (Sintonen, 1981), two with 15D.1 (Sintonen and Pekurinen, 1993; Sintonen, 1989) and two with the 15D.2 (Sintonen, 1994a and b). All four applications have used version I of the 15D (Table 3.6). These have been supplemented by four unpublished studies described in reviews of the instrument (Sintonen and Pekurinen 1993; Sintonen 1994a).

3.4.4.3 Practicality

This is an easy and brief questionnaire to use. Sintonen reports that it takes between 5 to 10 minutes to complete. He also reports the response rates to have been between 65-80% depending on whether reminders were used or not. In studies of hip and knee problems, the rates were 100% in hospital and 87% by post. Completion rates have been between 96-99%.

3.4.4.4 Reliability

In an unpublished study of patients waiting for coronary artery bypass grafts, the differences by dimensions between test and re-test at three months were found to be -.05 to 0.03, and none was significant. The percentages lying within two standard deviations of the mean difference were 92-100%, comparing favourably to NHP results on the same patients. Sintonen and Pekurinen (1993) also report that in a study of primary care

centre attenders scores at six months, there had been 'virtually no average change', though they did not present any details.

Most applications have used a self-administered version of the questionnaire, but Sintonen (1994a) has reported on a comparison between the responses of cancer patients and their personal nurses. Nurses were found to rate their patients as having significantly better health.

3.4.4.5 Descriptive validity

Content and face validity

The original 12D version was based on a review of official health policy documents published in Finland and was intended to cover the three areas identified by the WHO definition. The 15D incorporated advice from the medical profession, and Sintonen notes a particular concern with the apparent neglect of mental health in 12D. Dimensions for depression, distress and pain were added.

The largely 'expert' driven development was then followed by two surveys of primary care centre patients (n>2000). The respondents were asked to identify those aspects of health not included in the 15D, and their suggested additions were subsequently assigned by a researcher into four categories: clinical conditions, physical symptoms, vitality and mental problems. On the basis of these results, feedback from the uses of 15D.1 and an u-eported factor analysis, changes were made to the dimensions and their levels to form 15D.2. The number of levels was increased to five for all dimensions to improve sensitivity.

The 15D would appear to be very broad in its coverage compared with other QALY instruments. However, there has been no critical review of its content or the face validity.

Construct validity

There have been few published studies using the instrument. Sintonen (1994a) refers to some extreme group comparisons. It was found that the elderly (>65 year old) had a

lower score on every dimension of the 15D.2 (p=.001) than a younger group (17-35 years) except depression. People reporting an illness also had a lower mean score on all dimensions. In a cross-sectional study of patients before and after hip and knee replacements, post-operative patients were found to be significantly better in their mobility, work, social, pain and perceived health (Rissanen et al., 1995). Distinctive health profiles were also found for bypass and depression patients compared to the general population (Sintonen, 1994a).

Depression and distress scores of the 15D.1 were found to correlate with the Hamilton Depression Rating Scale (HMRS), a widely used condition-specific questionnaire, by -0.62 and -0.59. The scores on the 15D dimensions were able to predict correctly whether the HMRS score was more than 16 or not 77% of the time compared to 81% for the mental health dimension of the SF-36 (Sintonen, 1994a). The dimension scores of the 15D were also found to converge more with similar than dissimilar dimensions of the NHP and EQ-5D.

The sensitivity of the classification has been examined in terms of the percentages of respondents on the 'ceiling' and 'floor' of comparable dimensions. Sintonen (1994a) found the 15D to be the same or better in these terms than EQ -5D in a general population data set for all dimensions except mobility. This evidence suggests that the extra levels make it more sensitive than EQ-5D. It was found to have more in the top category in patients with depression than the SF-20, an earlier version of the SF-36, in mobility (74.9% vs. 25.6%), pain (21.8% vs. 14.4%) and social participation (21.8 vs. 12.6), but the same for mental health and slightly better in working (8.7% vs. 15.8%).

As a description of health, the 15D l shows promise. The large size of its classification makes it more sensitive than the EQ-5D, although the evidence is based on a very limited number of studies and range of conditions. The question is whether the large size of this measure presents any difficulties in valuation.

3.4.4.6 Valuation

The valuation of the 15D.2 has been based on a random sample of the Finnish population with useable response rate of around 30% (Sintonen, 1994b). There is evidence from the cross-country comparisons undertaken by the Euroqol Group that the values for hypothetical health states are similar between countries (Brook et al., 1991). However, poor response has an adverse effect on representativeness. There might also be concern about the quality of the data from a postal survey, but there were few inconsistencies found within dimensions.

The scale used to rate the relative importance of the dimensions was a cross between a visual analogue scale, as used by the EQ group, and magnitude estimation. In the instructions to respondents and in the way the scale is labelled, they are asked to regard it as a ratio scale: *'If, for example, an attribute is in your opinion half (1/2 or 50%) as important as the most important one, draw a line from the box following it to 50 on the scale'*. The same method was used to estimate the relative 'desirability' of dimension levels. This does not provide a valid cardinal measure of preferences. There was an attempt to estimate a utility function by transforming the ratings using the power relationships estimated by Torrance and his colleagues, but for reasons explained below, these functions were rejected for generating unlikely health state values.

The 15D.1 was valued using an additive formula that assumes the weight given to a dimension is unaltered by its level. This assumption was relaxed in the valuation of 15D.2 by re-estimating the weights for dimensions at the bottom of their level and these were found to be significantly different from those estimated with the levels set to the top. The intermediate levels of each dimension are assumed to be a linear extrapolation from the top and bottom level weights. This revised additive model is the one recommended by Sintonen (1994b). A multiplicative model was also estimated; however, the health state values predicted by the multiplicative models did not produce credible estimates. For example, according to this model 24.9% of the general population in Finland had a health state worse than death! This result was improved by replacing all negative valuations in the data set with 0.01, but then it was found that the model was very poor at distinguishing between states defined by the classification.

The 15D a decompositional approach was chosen because it would not have been possible to value directly 15 dimensional health states. However, there are concerns with the ability of this to predict health state values. Sintonen (1994b) found substantial differences between predicted values and those from respondents' ratings of their own states, but did not explore the data for any systematic differences.

### 3.4.4.7 Empirical validity

There have been no published applications of the 15D.2 and only a few for the 15D.1. In a cross-sectional study of patients waiting for hip and knee angioplasty, there were significant differences between the pre and post surgery groups (Rissanen et al., 1995). The prospective study of patients receiving hip and knee replacements found significant improvements six months after surgery. The average 15D score in coronary bypass candidates was also found significantly to improve between baseline and three months after the operation.

The study by Nord et al. (1993) found the 15D produced values of a similar magnitude to PTO (differences were -0.04 to 0.15) for four EQ-6D states. However, for reasons explained earlier this study had a number of serious methodological weaknesses.

### 3.4.4.8 Overview - key points

- 15D is a brief and easy-to-use self-completed questionnaire.
- There is some evidence of re-test reliability.
- It has a broad coverage of health domains
- There have been few studies using the instrument, but initial results are promising for its descriptive validity.
- There is no theoretical support for the ability of VAS values to reflect preferences on a cardinal scale, and a decompositional approach to estimating health state values must be tested.
- There is little evidence on the empirical validity of the 15D.

### 3.4.5 EQ-5D

This instrument was developed by a multidisciplinary group of researchers from seven centres across five countries (Euroqol Group, 1990). The original version had six dimensions, the EQ-6D, which has been succeeded by the five dimensional EQ-5D.

#### 3.4.5.1 Description

The five dimensions of the EQ-5D are mobility, self-care, usual activities, pain/discomfort and anxiety/depression. They each have three levels and together define 243 health states. Surveys to value samples of EQ-5D health states have been undertaken using a VAS rating scale (van Agt et al., 1994; Badia et al., 1995; and Selai and Rosser,1995). However, the most important valuation work with the EQ-5D has been a large-scale survey undertaken in the UK by the Measurement and Valuation of Health (MVH) group at York. Their work produced the TTO algorithm for valuing the EQ-5D presented in Appendix 1.5. It is an additive formula with decrements for the moderate and severe dysfunctional categories of the five dimensions, a constant term for any kind of dysfunction and the term 'N3' for whenever any of the dimensions are severe. Separate algorithms are available for different socio-demographic groups.

#### 3.4.5.2 Published literature

The search identified 40 publications, including refereed articles in journals, chapters of books, research reports, and conference papers (Table 3.2). The 'grey' literature has been particularly important for this instrument because the EQ-5D is a comparatively recent instrument, and much of the existing work has not been published. Twenty nine papers are concerned with methodology. There were eight studies using the EQ-5D, and this includes an MRC report and a conference paper, and one published application of the EQ-6D (Table 3.7). Two of the papers were found to be irrelevant for this review and so are not considered further[7].

[7] The trial of treatments of menorrhagia by Sculpher and colleagues (1993) did not use the descriptive part of EQ-5D. The study of Gastric cancer patients by Norum and Angelsen (1995) involved oncologists classifying and scoring the patients and so does not use the instrument in the recommended fashion.

3.4.5.3 Practicality

This is an easy-to-use and brief self-completed questionnaire of just two pages. It can be made simpler by using just the one page with the descriptive classification. By self-completion or interview administration it takes only a few minutes. The claim by Humphreys et al. (1995) that it 'usually' took 10 minutes does not seem reasonable.

Four out of the five studies reported response rates of more than 80% when the EQ-5D was being used to describe health alongside other, often lengthier, instruments. Studies of COPD and rheumatoid arthritis patients were able to achieve response rates in excess of 90%. Completion rates were over 90% in four out of five studies. No study reported any problems in getting patients to complete this instrument.

3.4.5.4 Reliability

Three studies have examined the re-test reliability of the EQ-5D; one in a sample of elderly women aged 75 or over, the second in a sample of patients with COPD attending a chest clinic and the third a longitudinal study of patients with rheumatoid arthritis (Brazier et al., 1996a, 1996b; Hurst 1996). In the first two, the correlations between the test and re-test single index scores (based on an interim algorithm) in patients who said their health had not changed after an interval of 6 months were 0.67 and 0.83 respectively. The mean difference was non-significant and within a 95% confidence interval of plus or minus 0.05. The reliability coefficient in the RA patients was 0.55. In all studies, these results compared well with the other generic and condition-specific health measures.

3.4.5.5 Descriptive validity

Content and face validity

The original instrument was developed from a review of other health status measures, including the QWB, Sickness Impact Profile, Nottingham Health Profile and the Rosser classification (The Euroqol Group, 1990). Kind (1996) has described the process as one

where *'...researchers principally drew on their own expertise and the evidence available from the literature in order to determine the dimensions of interest'*. The aim was to develop an instrument which addressed a 'core' of domains common to other generic health status questionnaires and which reflected the most important concerns of patients themselves. It is not intended to cover all aspects of health and is inevitably the result of a compromise between being comprehensive and the need to keep the instrument simple enough for the chosen valuation strategy, namely the valuation of entire health states (Williams, 1995).

On the basis of experience gained from using this instrument the group developed the EQ-5D. The number of dimensions was reduced to 5 by combining family/ leisure activity with main activity to form 'usual' activity. This it has been argued was justified on the grounds that social relations were found to contribute little to health state valuations, though no evidence has been brought forward to support this claim (Kind, 1996). The number of levels was raised to three for all dimensions in order to achieve *'a more balanced structure for each dimension, giving equal salience to each component in the resulting composite health state'* (Kind, 1996). The group did not include a dimension for energy since it was found to have no impact on health state valuations (Bjork, 1991).

The MVH group at York have conducted a survey in the West Midlands to assess the coverage of the content validity of EQ-5D and other measures of health (Rosser, NHP, QWB and SIP) i.e. to establish *'..what the general population regard as the salient feature of health'* (Williams, 1995). The survey recruited samples of the general population for interview (young disabled and carers of disabled children were also interviewed). An unprompted section of the interview asked individuals to list the distinguishing features of 'good' and 'bad' health. The results for the general population sample (n=196) was a list of 20 items covering activities, feelings, symptoms, and general well-being. The five most commonly mentioned health domains were feelings, energy, usual activities, appearance and mobility with a total coverage of 45%. The items varied little in importance according to the respondents. Energy, sleep, visual acuity, hearing and many symptoms of diseases, were excluded from the EQ-5D. The

EQ-5D was found to cover 35.9% of the health items mentioned by individuals in the unprompted section, compared to 26.9% for the Rosser, 49.1% for SIP, 58.6% for NHP and 58.6% for QWB.

The face validity of the EQ-5D has been criticised for having only three categories per dimension, which are thought to be too insensitive for detecting smaller changes (McDowell and Newell, 1996). A high proportion of people have been found on the ceiling of the classification i.e. recording no problem (Brazier et al., 1993; Hollingworth et al., 1995). In a general population survey using EQ-6D, there were 95% or more of respondents in the top category of mobility, self care, main activities and family/leisure, indicating no problems, compared to 37-72% for the SF-36 (Brazier et al., 1993). EQ-5D has slightly more categories and could be less prone to skewness. The national MVH survey using the EQ-5D found the number at the top of the mobility dimension was reduced to 88.6% and to 86.3 for usual activities.

Construct validity

In the general population survey by Brazier and colleagues (1993), patients who responded as having no health problem on dimensions of the EQ-6D were sub-divided into those who had at least the median SF-36 score (better health) and those who scored less than the median on comparable dimensions (worse health). Patients in the poor health groups were found to have a higher mean age, a higher proportion of women and a higher proportion of patients not in full-time employment than the better group. The poor groups were also more likely to have consulted a GP recently, attended outpatients in the last three months, or been an inpatient in the last year. This evidence suggests the EQ-6D classification is less sensitive at detecting perceived health problem than the SF-36.

Two studies have examined the validity of the dimensions of the EQ-5D. Patients diagnosed with migraine were found to be significantly worse than a general population sample in terms of pain, anxiety and depression and usual activities (Essink-Bot et al., 1995). Hollingworth and his co-workers (1995) studied a group of patients referred for an magnetic resonance scan (MRI) of the knee. The EQ-5D was able to show these

patient groups to be significantly worse on its unscored dimensions. Four other studies have examined the sensitivity of the index. It has been shown to distinguish between COPD patients and the general population (Harper et al., 1997) and migraineurs and the general population (Essink-Bot et al., 1995). Furthermore, the EQ-5D index has been able to detect differences within disease groups in patients with COPD (severe vs. not severe as defined by the Fletcher scale) and rheumatoid arthritis patients by functional class (Hurst, 1996). However, it was not able to distinguish significantly between COPD groups defined in terms of a 6 minute walk test nor on the basis of whether or not they had a comorbidity, in contrast to several dimensions of SF-36 (Harper et al., 1995).

The EQ-5D index has been found to correlate moderately well with other generic and condition-specific measures (Brazier et al., 1993; Hurst et al., 1994). It has also been shown to reflect changes in the health. The EQ-5D score improved in patients who had been for a knee scan over a six month period (Hollingworth et al., 1995), before and after reconstruction in vascular disease patients (Humphreys et al., 1995) and in patients who reported a change in their rheumatoid arthritis (Hurst, 1996).

3.4.5.6 Valuation

The MVH survey was based on a large sample (n=3395), broadly representative of the UK population (in terms of a range of socio-demographic, health and health service use variables), and achieved a response rate of 64% (higher than previous valuation surveys using the EQ-5D). Interviews were conducted by trained staff using well designed and tested visual aids (Thomas R, Thomson K, 1992; Dolan et al., 1996). The quality of data in terms of completeness and consistency was impressive and has been well documented (MVH, 1994).

The TTO technique has considerable support amongst many health economists as a measure of preferences. The statistical modelling to estimate health states values used random effects to allow for between respondent variation and examined alternative specifications (including interaction effects). A simple additive model was chosen on grounds of its goodness fit of the data ($R^2$ of 0.46) and parsimony compared to other specifications. The model contains decrements for each of the moderate and severe

dysfunctional categories of the five dimensions, a constant for any kind of dysfunction and the term 'N3' for whenever any of the dimensions are severe. The model suffered from heteroscedasticity and failed a test of specification, but the authors claimed this was unavoidable with such a large data set and found it did not harm the robustness of the estimates (which were confirmed in a split sample test).

### 3.4.5.7 Empirical validity

The results of the MVH survey only became available to researchers from the beginning of 1996, and there are no published studies using the new tariffs. Until recently researchers have been using a scoring system based on a simpler model estimated by ordinary least squares regression, known as the interim tariff (personal communication, June 1994).

The single index derived from the EQ-5D using the interim tariff has been found to distinguish between the general population patients and COPD (Harper et al., 1995), Migraineurs (Essink-Bot et al., 1995) and those awaiting an MRI scan of the knee (Hollingworth et al., 1995). The detection of differences within disease group in patients with COPD (severe vs. not severe as defined by the Fletcher scale) and rheumatoid arthritis patients (functional class) is also in line with expectations. It has also been shown to reflect hypothesised changes in the health. The EQ-5D score improved in patients who had been for a knee scan over a six month period, before and after reconstruction in patients with vascular disease patients and in patients who reported a change in their rheumatoid arthritis.

The EQ-5D index was not able to detect a significant change in COPD patients who said their health had changed between assessments, despite statistically significant changes in dimensions of SF-36 and the condition-specific measures (Brazier et al., 1995). In knee patients followed up after an MRI scan, the group reporting no change according to the EQ-5D index were, however, found to have changed according to the SF-36 (Hollingworth et al., 1995). Evidence from this second study was not supported by any other indicator of change and hence must be treated with some scepticism.

3.4.5.8 Overview - key points

- It is a very brief and easy-to-use instrument.
- There is evidence of its re-test reliability.
- The dimensions cover many though not all domains of health. The three levels would on the face of it seem too crude to detect smaller changes.
- There is little evidence on construct validity, but what is available suggests it can detect large differences, though there is some evidence of insensitivity.
- TTO is an accepted method for deriving preference values and the MVH survey in the UK is impressive and the statistical modelling rigorous.
- Crude comparisons show EQ-5D is able to detect large differences in line with expected preferences, though there is some contrary evidence against patient perceived health.

## 3.5 Discussion and conclusion

All measures use a short list of questions to be administered by self-completion or interview in less than 10 minutes, with the exception of the QWB. The QWB has a lengthier interview schedule, which involves detailed probing of the respondents which can take 20 minutes. All instruments were able to achieve high levels of response and completion and there was little to choose between the questionnaires on the basis of practicality.

Evidence has been found of differences between the assessment by patients of their own health compared to that of health professionals using the Rosser and HUI. This implies that the method of administering these instruments must be standardised. There is evidence of re-test reliability for EQ-5D and 15-D, but this property has not been adequately investigated in any of the five measures.

The descriptive content of the measures differ widely. The size varies between the Rosser, with just two dimensions, compared to the 15 dimensions of the 15D. All measures cover physical functioning, though there are differences in whether the concept is described in terms of capacity (e.g. HUI) or actual behaviour and

performance (e.g. QWB). The coverage of symptoms, mental health and social health is less consistent. The QWB explicitly excludes mental health as a separate dimension, but has a long list of symptoms and problems. The HUI-III covers many of the symptoms or health problems, but does not examine role or social function, since these are regarded as 'out of skin' and not appropriate in a measure of individual health preferences. The EQ-5D has dimensions for role and social function, and pain and mood, but not for many other symptoms and health problems.

In terms of content no single measure dominates. The Rosser seems to be inferior to the others in terms of its coverage. The choice from the remaining four will depend on what aspects of health the potential user wishes to cover. Despite the claim that these are generic measures, they do not cover the exactly the same aspects of health. Their relevance may therefore vary depending on the disease group and by age of the patients being evaluated. The HUI measures (particularly HUI-II) may be better suited to a younger population than the EQ-5D, for example, though this has not been tested. There are also issues about perspective and whether or not social health is relevant.

Preference-based measures have been criticised for being crude and insensitive. However, there was evidence for all measures of their ability to detect differences in group comparisons and the scores were significantly correlated with other measures of the health. It is difficult to compare the performance of the measures owing to differences in the quantity and type of evidence available on each measure. Most of the evidence on the QWB scale was limited to correlations with related health status measures, with very little detailed scrutiny of the descriptive classification, whereas evidence for the HUI-I was limited to survivors of childhood cancer. There was some suggestion of insensitivity in all measures, except the 15D where there have been too few studies.

The QWB, Rosser, and the 15D can be regarded as inferior to the other two measures owing to their use of VAS and ME to value the health descriptions. HUI-II and III might be preferred to the EQ-5D by those who regard the SG as the 'gold standard' (see Chapter 5). However, the values have been derived from VAS on the basis of a power

function which can be criticised on theoretical and empirical grounds. The valuation of the HUI has been based on a smaller and less representative sample of the general population than the MVH survey. The virtues of the algebraic approach used by HUI versus statistical methods used to value the EQ-5D has not been addressed in the literature. However, there is evidence to suggest the algebraic method may be poor at predicting health state values.

Evidence on empirical validity has been very limited. The QWB has been shown to correlate with direct preference elicitation, but such evidence has not been published for the EQ-5D and HUI-I. There is evidence of the EQ-5D converging with patient perception of health change in one study but not another. There was no evidence found on the correlation of the HUIs with stated preferences. The measures were found to reflect hypothesised preferences between patient groups, but the evidence would appear too limited to draw firm conclusions.

This review concludes that the best preference-based measures at the moment would seem to be the EQ-5D and the HUIs. For the HUI there is a further choice between versions. The HUI-I would seem to have many problems and has not been used by its developers for many years. The HUI-II is designed for children and HUI-III has been designed for adults. At the time of writing, there are no published weights available for HUI-II. For economic evaluations alongside clinical trials of interventions for adults the EQ-5D is preferred. This conclusion would have to be re-appraised when (Canadian) weights become available for the HUI-III. However, in the UK, researchers are likely to continue to favour the EQ-5D on the grounds that it has been more widely used this country and there are UK weights.

This review has served to place the empirical research of this thesis into context. It has identified the marked lack of evidence about the reliability and validity of existing preference-based measures of health, particularly their sensitivity to detect the small difference likely to emerge in trials comparing alternative treatments for the same condition. The review identified some evidence of insensitivity in the classification of the EQ-5D. The HUI-II and III are considerably larger than the EQ-5D and hence

potentially are more sensitive, but they cover different aspects of health and there is no evidence of whether these are more sensitive than the EQ-5D. Furthermore, they use the algebraic method to estimate the weights and there are doubts about the predictive validity of this approach.

There would seem to be a case for examining the potential for developing a larger and more sensitive preference-based measure of health than the EQ-5D. The questions addressed by the research in this thesis are whether a larger classification, based on the SF-36, would be able describe finer differences between states that are important in terms of preferences, and given the concerns identified in the review about the algebraic methods used to value the HUIs, whether statistical methods can be used to value such a large classification. Before embarking on this research, it is necessary to examine critically the SF-36 and assess its potential for providing the basis for a new preference-based measure.

**Table 3.1: Search strategy**

| | |
|---|---|
| **Rosser* classification** | **Quality adjusted life year*** |
| **Rosser matrix** | **QALY*** |
| **Rosser distress {categor*/state*}** | **Classification of illness states** |
| **Health Measurement Questionnaire** | **15D** |
| **Index of health-related quality of life** | **15 dimension*** |
| **Index of wellbeing** | **12D** |
| **Index of well-being** | **12 dimension*** |
| **Quality of wellbeing** | **Euroqol** |
| **Quality of well-being** | **Euroqolc** |
| **QWB** | **Well year*** |
| **Health utilities ind*** | **Multiattribute* utilit*** |
| **Heath states utility ind*** | **Multi attribute* utilit*** |
| **Multiattribute* health ind*** | **Multi attribute* health state*** |
| **Multi attribute* health ind*** | **Multiattribute* health state*** |
| **Multi attribute* theor*** | **Multi attribute* theor*** |
| **Multiattribute* analys*** | **Multi attribute* analys*** |
| **HUI** | |

**Table 3.2: Papers identified for review**

**QWB**

Anderson, G.M. (1982).
Anderson, J.P., Bush, J.W. and Berry, C.C. (1986)
Anderson, J.P., Bush, J.W. and Berry, C.C. (1988)
Anderson, J.P., Kaplan, R.M., Berry, C.C., Bush, J.W. and Rumbaut, R.G. (1989)
Anderson, J.P., Kaplan, R.M. and Schneiderman, L.J. (1994)
Andresen, E.M., Patrick, D.L., Carter, W.B. and Malmgren, J.A. (1995)
Bakker, C.H., Rutten van Molken, M., van Doorslaer, E., Bennett, K. and van der Linden, S. (1993)
Balaban, D.J., Sagi, P.C., Goldfarb, N.I. and Nettler, S. (1986)
Bombardier, C. and Raboud, J. (1991)
Bombardier, C., Ware, J., Russell, I., Larson, M.G., Chalmers, A. and Leighton Read, J. (1986)
Bradlyn, A.S., Harris, C.V., Warner, J.E., Ritchey, A.K. and Zaboy, K. (1993)
Bush, J.W., Anderson, J.P., Kaplan, R.M. and Blischke, W.R. (1982)
Calfas, K.J., Kaplan, R.M. and Ingram, R.E. (1992)
Carr-Hill, R.A. and Morris, J. (1991)
de Groot, J., de Groot, W., Kamphuis, M., Vos, P.F., Berend, K. and Blankestijn, P.J. (1994)
Dirksen, S.R. (1995) Search for meaning in long-term cancer survivors. *J. Adv. Nurs.* **21**, 628-633.
Elvik, R. (1995)
Erickson, P., Kendall, E.A., Anderson, J.P. and Kaplan, R.M. (1989)
Fryback, D.G., Dasbach, E.D., Klein, R., Klein, B.E.K., Martin, P.A., Dorn, N. and Peterson, K. (1992)
Fryback, D.G., Dasbach, E.J., Klein, R., Klein, B.E., Dorn, N., Peterson, K. and Martin, P.A. (1993)
Ganiats, T.G., Palinkas, L.A. and Kaplan, R.M. (1992)
Gilbert, A., Owen, N., Innes, J.M. and Sansom, L. (1993)
Holbrook, T.L., Hoyt, D.B., Anderson, J.P., Hollingsworth-Fridlund, P. and Shackford, S.R. (1994)
Hornberger, J.C., Redelmeier, D.A. and Petersen, J. (1992)
Kaplan, R.I. and Atkins, C.J. (1989)
Kaplan, R.M. (1989)
Kaplan, R.M. (1993a)
Kaplan, R.M. (1993b)
Kaplan, R.M. (1994a)
Kaplan, R.M. (1994b)
Kaplan, R.M. and Anderson, J.P. (1988)
Kaplan, R.M., Anderson, J.P., Patterson, T.L., McCutchan, J.A., Weinrich, J.D., Heaton, R.K., Atkinson, J.H., Thal, L., Chandler, J. and Grant, I. (1995)
Kaplan, R.M., Anderson, J.P. and Wingard, D.L. (1991)
Kaplan, R.M., Anderson, J.P., Wu, A.W., Mathews, W.C., Kozin, F. and Orenstein, D. (1989)
Kaplan, R.M., Atkins, C.J. and Timms, R. (1984)
Kaplan, R.M. and Bush, J.W. (1982)

Kaplan, R.M., Bush, J.W. and Berry, C.C. (1976)
Kaplan, R.M., Bush, J.W. and Berry, C.C. (1979)
Kaplan, R.M., Coons, S.J. and Anderson, J.P. (1992)
Kaplan, R.M., Debon, M. and Anderson, B.F. (1991)
Liang, M.H., Fossel, A.H. and Larson, M.G. (1990)
Manzetti, J.D., Hoffman, L.A., Sereika, S.M., Sciurba, F.C. and Griffith, B.P. (1994)
Mold, J.W., Holtgrave, D.R., Bisonni, R.S., Marley, D.S., Wright, R.A. and Spann, S.J. (1992)
Nord, E. (1993)
Orenstein, D.M. and Kaplan, R.M. (1991)
Orenstein, D.M., Nixon, P.A., Ross, E.A. and Kaplan, R.M. (1989)
Orenstein, D.M., Pattishall, E.N., Nixon, P.A., Ross, E.A. and Kaplan, R.M. (1990)
Patrick, D.L., Bush, J.W. and Chen, M.M. (1973)
Patrick, D.L., Bush, J.W. and Chen, M.M. (1973)
Read, J.L., Quinn, R.J. and Hoefer, M.A. (1987)
Reed, P.G. (1986)
Schneiderman, L.J., Kronick, R., Kaplan, R.M., Anderson, J.P. and Langer, R.D. (1992)
Tandon, P.K., Stander, H. and Schwarz, R.P., Jr. (1989)
Tramarin, A., Milocchi, F., Tolley, K., Vaglia, A., Marcolini, F., Manfrin, V. and de-Lalla, F. (1992)
Visser, M.C., Fletcher, A.E., Parr, G., Simpson, A. and Bulpitt, C.J. (1994)
Wu AW, Mathews WC, Brysk LT, Hampton Atkinson J, Grant I, Abramson I, Kennedy CJ, McCutchan JA, Spector SA and Richman DD (1990)

**Rosser**

Bryan, S., Parkin, D. and Donaldson, C. (1991)
Carr-Hill, R.A. and Morris, J. (1991)
Chan, C.L.H. and Villar, R.N. (1996)
Coast, J. (1992).
Cole, R.P., Shakespeare, V., Shakespeare, P. and Hobby, J.A. (1994)
Donaldson, C., Atkinson, A., Bond, J. and Wright, K. (1988a)
Donaldson, C., Atkinson, A., Bond, J. and Wright, K. (1988b)
Elvik, R. (1995)
Gater, R.A., Kind, P. and Gudex, C. (1995).
Glasziou, P.P., Bromwich, S. and Simes, R.J. (1994)
Gudex, C. (1986).
Gudex, C. and Kind, P. (1988)
Gudex, C. and Kind, P. (1991)
Gudex, C., Kind, P., van Dalen, H., Durand M-A, Morris, J. and Williams, A. (1993)
Gudex, C., Williams, A., Jourdan, M., Mason, R., Maynard, J., O'Flynn, R. and Rendall, M. (1990)
Gudex, C.M. (1995)
Hollingworth, W., Mackenzie, R., Todd, C.J. and Dixon, A.K. (1995)
Humphreys, W.V., Evans, F., Watkin, G. and Williams, T. (1995)
Kallis, P., Unsworth White, J., Munsch, C., Gallivan, S., Smith, E.E., Parker, D.J., Pepper, J.R. and Treasure, T. (1993)
Kind, P. (1990)
Kind, P. and Gudex, C.M. (1994)

Kind, P. and Rosser, R. (1988)
Kind, P., van Dalen, H., Morris, J. and Williams, A. (1993)
Launois, R., Henry, B., Marty, J.R., Gersberg, M., Lassale, C., Benoist, M. and Goehrs, J.M. (1994)
Lonnqvist, J., Sihvo, S., Syvalahti, E., Sintonen, H., Kiviruusu, O. and Pitkanen, H. (1995)
Mackenzie, R., Hollingworth, W. and Dixon, A.K. (1994)
Magee, T.R., Scott, D.J., Dunkley, A., St Johnston, J., Campbell, W.B., Baird, R.N. and Horrocks, M. (1992)
Normantaylor, F.H., Palmer, C.R. and Villar, R.N. (1996)
Payne, S.P. and Galland, R.B. (1995)
Petrou, S., Davey, P. and Malek, M. (1992)
Rabin, R., Rosser, R.M. and Butler, C. (1993)
Rawles, J., Light, J. and Watt, M. (1992)
Rosser, R., Allison, R., Butler, C., Cottee, M., Rabin, R. and Selai, C. (1993)
Rosser, R.M. and Kind, P. (1978)
Rosser, R.M. and Watts, V.C. (1972)
Unsworthwhite, J., Kallis, P., Treasure, T. and Pepper, J.R. (1994)
van Dalen, H., Williams, A. and Gudex, C. (1994)
Wade, D.T. (1991)
Watkins, L.D., Bell, B.A., Marsh, H.T. and Uttley, D. (1990)
Whynes, D.K. and Neilson, A.R. (1993)
Whynes, D.K., Neilson, A.R., Robinson, M.H. and Hardcastle, J.D. (1994)
Williams, A. (1985)

**HUI**

Barr, R.D., Feeny, D., Furlong, W., Weitzman, S. and Torrance, G.W. (1995)
Barr, R.D., Furlong, W., Dawson, S., Whitton, A.C., Strautmanis, I., Pai, M., Feeny, D. and Torrance, G.W. (1993)
Barr, R.D., Pai, M.K.R., Weitzman, S., Feeny, D., Furlong, W., Rosenbaum, P. and Torrance, G.W. (1994)
Boyle, M.H., Furlong, W., Feeny, D., Torrance, G.W. and Hatcher, J. (1995)
Boyle, M.H. and Torrance, G.W. (1984)
Boyle, M.H., Torrance, G.W., Sinclair, J.C. and Horwood, S.P. (1983)
Cadman, D. and Goldsmith, C. (1986)
Cadman, D., Goldsmith, C. and Bashim, P. (1984)
de Groot, J., de Groot, W., Kamphuis, M., Vos, P.F., Berend, K. and Blankestijn, P.J. (1994)
Elvik, R. (1995)
Erickson, P., Kendall, E.A., Anderson, J.P. and Kaplan, R.M. (1989)
Feeny, D., Furlong, W., Barr, R.D., Torrance, G.W., Rosenbaum, P. and Weitzman, S. (1992)
Feeny, D., Furlong, W., Boyle, M. and Torrance, G.W. (1995)
Feeny, D., Leiper, A., Barr, R.D., Furlong, W., Torrance, G.W., Rosenbaum, P. and Furlong, W., Torrance, G.W. and Feeny, D. (1993)
Gold, M., Franks, P. and Erickson, P. (1996)
Saigal, S., Feeny, D., Furlong, W., Rosenbaum, P., Burrows, E. and Torrance, G. (1994)
Saigal, S., Rosenbaum, P.L., Furlong, W.J., Feeny, D.H. and Burrows, E. (1995)

Torrance, G.W., Boyle, M.H. and Horwood, S.P. (1982)
Torrance, G.W., Furlong, W., Feeny, D. and Boyle, M. (1995)
Verhoef, C.G., Verbeek, A.L., Stalpers, L.J. and van Daal, W.A. (1990)

**15D**

Apajasalo, M., Sintonen, H., Holmberg, C., Sinkkonen, J., Aalberg, V., Pihko, H., Siimes, M.A., Kaitila, I., Makela, A., Rantakari, K., Anttila, R. and Rautonen, J. (1996)
Lonnqvist, J., Sihvo, S., Syvalahti, E., Sintonen, H., Kiviruusu, O. and Pitkanen, H. (1995)
Lonnqvist, J., Sintonen, H., Syvalahti, E., Appelberg, B., Koskinen, T., Mannikko, T., Mehtonen, O.P., Naarala, M., Sihvo, S., Auvinen, J. and et al (1994)
Rissanen, P., Aro, S., Sintonen, H., Slatis, P. and Paavolainen, P. (1996)
Rissanen, P., Aro, S., Slatis, P., Sintonen, H. and Paavolainen, P. (1995)
Sintonen, H. (1981)
Sintonen, H. (1993)

**EuroQol**

Anderson, R.T., Aaronson, N.K. and Wilkin, D. (1993)
Brazier, J., Jones, N. and Kind, P. (1993)
Brazier, J., Walters, S.J., Nicholl, J.P. and Kohler, B. (1996a)
Brooks RG, Jendteg S, Lindgren B, Persson U and Bjork S (1991)
Caperna, J. and Mathews, W.C. (1996)
Carr-hill, R.A. (1991)
Carr-hill, R.A. (1992)
Dolan, P. (1994)
Dolan, P., Gudex, C., Kind, P. and Williams, A. (1995)
Dolan, P., Gudex, C., Kind, P. and Williams, A. (1996)
Elvik, R. (1995)
Essink-Bot, M.L., Bonsel, G.J. and Van Der Maas, P.J. (1990)
Essink-Bot, M.L., Stouthard, M.E. and Bonsel, G.J. (1993)
Essink-Bot, M.L., Vanroyen, L., Krabbe, P., Bonsel, G.J. and Rutten, F.F.H. (1995)
Euroqol group (1990)
Euroqol group (1991)
Euroqol group (1992).
Gravelle, H. (1995)
Hollingworth, W., Mackenzie, R., Todd, C.J. and Dixon, A.K. (1995)
Hurst, N.P., Jobanputra, P., Hunter, M., Lambert, M., Lochhead, A. and Brown, H. (1994)
Kind, P. (1994)
Kind, P. (1996)
Kind, P., Gudex, C., Dolan, P. and Williams, A. (1994)
MVH group (1994)
MVH group (1995)
Nord, E. (1991a)
Nord, E. (1991b)
O'Hanlon, M., Fox Rushby, J. and Buxton, M.J. (1994)
Parkin, D. (1991)
Rosser, R. and Sintonen, H. (1993)

Sculpher, M., Bryan, S., Dwyer, N., Hutton, J. and Stirrat, G.M. (1993)
Selai, C. and Rosser, R. (1995)
Thomas, R. and Thomson, K. (1992)
van Agt, H.M., Essink-Bot, M.L., Krabbe, P.F. and Bonsel, G.J. (1994)
van dalen, H., Williams, A. and Gudex, C. (1994)
Williams, A. (1993)
Williams, A. (1995)

**Table 3.3: Studies using the QWB**

| Study | Patient group | | Practicality | | | Reliability | Descriptive validity | | Empirical validity |
|---|---|---|---|---|---|---|---|---|---|
| | | n | timing | response rate | completion rate | | content and face | construct | |
| Anderson et al. 1989 | 5 patient groups | 1866 | - | - | - | yes[8] | yes | - | - |
| Anderson et al. 1994 | terminally ill | 204 | - | - | - | - | - | - | - |
| Andresen et al. 1995 | older adults (mean age 72.5) | 200 | 17.4 mins | 68% | 93% | - | - | yes | - |
| Bombardier et al. 1986,1991 | rheumatoid arthritis | 303 | 20 mins | - | - | - | yes | yes | - |
| Calfas et al. 1992 | osteoarthritis | 40 | - | - | - | - | - | yes | - |
| Fryback et al.l 1993 | random sample of healthy adults | 1356 | - | 86% | - | - | - | yes | yes |
| Ganaits et al. 1992 | patients with atrial fibrillation | 664 | - | - | - | - | - | yes | - |
| Ganiats et al. 1991 | neonatal circumcision | no data | - | - | - | - | - | - | - |
| Gilbert et al. 1993 | elderly care | 69 | QWB not used | | | | | | |
| Holbrook et al. 1994 | trauma | 61 | - | - | - | - | yes | yes | yes |
| Hornberger et al. | chronic renal failure | 83 | <10 mins. | 100%? | 100%? | - | - | yes | yes |
| Kaplan et al. 1976 | random sample of the general pop. | 867 | - | - | - | - | - | yes | yes |
| Kaplan et al. 1984 | COPD | 75 | - | 100%? | 100%? | - | - | yes | - |
| Kaplan et al. 1989 | arthritis | 83 | - | - | - | - | - | yes | - |
| Kaplan et al. 1995 | HIV | 514 | - | - | - | - | - | yes | yes |
| Liang et al. 1994 | arthritis patients | 50 | - | - | 98% | - | - | yes | yes |
| Manzetti et al. 1994 | lung transplants | 9 | - | - | - | - | - | - | - |
| Mold et al. 1992 | prostate screening | | Excluded - did not collect data | | | | | | |

[8] Yes indicates the study reports evidence on this critieria (for or against)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Orenstein et al. 1989 | cystic fibrosis | 44 | - | - | - | - | - | yes | - |
| Orenstein et al. 1990 | cystic fibrosis | 28 | - | - | - | - | yes | yes | - |
| Orenstein et al. 1991 | | | excluded - no data | | | | | | |
| Read et al. | various O/P & I/P | 400 | 18.2 mins | - | - | - | yes | - | - |
| Tandon et al. 1989 | congestive heart failure | 111 | - | - | - | - | yes | yes | - |
| Tramarin et al. 1992 | AIDS | 42 | - | - | - | - | - | - | - |
| Wu et al. 1990 | AIDS | 31 | 10 mins | - | - | - | - | yes | - |

**Table 3.4: Studies using the Rosser Classification of illness**

| Study | Patient group | | | Practicality | | | Relibialility | Descriptive validity | | Empirical validity |
|---|---|---|---|---|---|---|---|---|---|---|
| | | n | Method | timing | response rate | completion rate | | content and face | construct | |
| Bryan S et al. 1991 | chiropody | 84 | HMQ | - | - | - | yes | yes | yes | - |
| Chan & Villar 1996 | hip replacement | 176 | Mapping | - | - | - | - | - | yes | - |
| Coast 1992 | various | - | Mapping | - | - | - | - | yes | - | - |
| Cole et al. 1994 | plastic surgery | 292 | HMQ only | - | 73% | - | - | - | yes | - |
| Donaldson et al. 1988a&b | elderly care | | Interview | - | - | - | - | yes | yes | - |
| Drewett et al. 1992 | knee replacement | 26 | Mapping | - | - | - | - | yes | - | yes |
| Gater et al. 1995 | psychiatric care | 138 | HMQ only | | - | - | - | - | yes | - |
| Glasziou et al. 1994 | thrombolytic therapy | 776 | HMQ | - | 92% | - | - | - | yes | yes |
| Gudex 1995 | ESRF | 900 | HMQ | - | 78% | - | - | - | - | yes |
| Gudex et al. 1990 | various | - | Clinician | - | - | - | - | - | - | - |
| Hollingworth et al. 1996 | knee problems | 82 | HMQ | - | 84% at baseline | 87% at baseline | | - | yes | yes |
| Kallis et al. 1993; Unsworth-white et al. 1994 | cardiac surgery | 207 | HMQ | - | 95% | - | - | - | yes | yes |
| Kerridge 1995 | ICU | 132 | HMQ | - | - | - | - | - | - | - |
| Kind and Gudex 1994 | random sample of adults | 430 | HMQ by interview | - | 53% agreed to interview | 95.5% | - | - | yes | - |
| Launois et al.l 1994 | low back pain | 146 | HMQ | - | - | - | - | - | yes | - |
| Magee et al. 1992 | abdominal aortic aneurysm | 165 | HMQ Interview | 30 mins | - | - | - | - | yes | yes |
| Payne & Galland 1995 | aortic reconstruction | 93 | HMQ by interview | - | - | - | - | - | - | - |
| Petrou et al. 1992 | hip & knee replacement | 44& 159 | Self-assesent & observer assessment | - | - | - | yes | yes | - | yes |
| Rawles et al. 1992 | myocardial | 206 | Interview | - | - | - | - | - | yes | yes |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | infarction | | | | | | | | | |
| Watkins et al. 1990 | neurosurgery | 50 | Clinician assessment | - | - | - | - | - | - | - |
| Whynes & Neilson 1993 | colorectal cancer | 221 | HMQ & clinician assessment | - | - | - | yes | - | yes | - |
| Whynes et al. 1994 | colorectal cancer | 351 | HMQ | - | 76-85% | - | - | - | yes | - |
| Williams et al. 1985 | CABG | - | Clinician | - | - | - | - | yes | - | - |

**Table 3.5: Studies using the HUI[1]**

| Study | Patient group | | HUI version | Practicality | | | Reliability | | Descriptive validity | | Empirical Validity |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | n | | timing | response rate | completion rate | inter-rater | retest | content and face | construct | hypothetical preferences |
| Barr et al. 1994 | survivors of therapy for brain tumours | 10 | II by profs. & parents | - | 100% | 100% | yes | - | - | yes | - |
| Barr et al. 1993 | survivors of acute lymphoblastic leukemia (LL) | 55 | II by profs. | 1 min. | - (100%?) | - (100%?) | yes | - | - | yes | - |
| Billson and Walker 1994 | survivors of cancer | 63 | II by profs., parents & child | Doctors 2 mins & patients 5 mins | 79% | 96% | yes | - | - | - | - |
| Boyle et al. 1994 | general population | 555 | III by tele. inter. | - | 91.2% | | - | yes | - | - | - |
| Boyle et al. 1983 | low birth weight babies | | I by home inter. | - | - | - | - | - | - | yes | yes |
| Feeny et al. 1993a | childhood cancer | 28 | II - profs. | - | 100% | 100% | yes | - | - | yes | - |
| Feeny et al. 1993b | high risk acute LL | 69 | II by mapping | - | - | - | - | - | - | yes | - |
| Gold et al. 1996 | general population | >10 000 | I by mapping | - | - | - | - | - | - | - | yes |
| Kanabar et al. 1995 | survivors of cancer | 30 | modified postal II | - | 93% | 100% | - | - | - | - | - |
| Saigal et al. | low birth-weight children | 156 | II by mapping | - | - | - | - | - | - | yes | - |

**Table 3.6: Studies using the 15D**

| Study | Patient group | | Practicality | | | Reliability | Descriptive validity | | Empirical validity |
|---|---|---|---|---|---|---|---|---|---|
| | | n | timing | response rate | completion rate | | content and face | construct | |
| Lonnqvist et al. 1994 | Paients with depression | 209 | - | | 96% response and completion combined | - | - | - | Yes |
| Lonnqvist et al. 1995 | Patients with depression | 59 | - | - | - | - | - | - | - |
| Rissanen et al. 1995 | Hip and knee patients | 355 | - | 100% in hospital; 87% from post | - | - | - | Yes | Yes |
| Rissanen et al. 1996 | Hip and knee patients | 452 | - | | 79.5% returned & completed at 2 year follow-up | | | | |
| Unpublished studies | | | | | | | | | |
| Brommel 1990 | CABG | 93 | - | - | - | Yes | - | - | Yes |
| In: Sintonen 1995 | Cancer patients | 70 | - | - | - | Yes | | | |
| Pekurinen et al. 1991 | Attenders at primary care centres | 1815 | - | 72% | - | Yes | - | yes | Yes |
| In: Sintonen 1995 | Valuation samples for 15D (1&2) | 2007 | - | | Completion 96-99% | | | | |
| In: Sintonen 1995 | Random samples of general population | 500 | - | 72% | Completion 96-99% | - | - | yes | - |

**Table 3.7: Studies using the EQ-5D or 6D[1]**

| Study | Patient group | | Practicality | | | Reliability | Descriptive validity | | Valuations |
|---|---|---|---|---|---|---|---|---|---|
| | | n | timing | response rate | completion rate | retest reliability | content and face | construct | hypothetical preferences |
| Brazier et al. 1996a | Elderly (>75) | 380 | - | 99% | >90% | yes | - | | yes |
| Brazier et al. 1993 | General population (16-74) | 1980 | - | 83% | >95% | - | - | yes | yes |
| Harper et al. 1997 | COPD | 142 | - | 91% | 92% | yes | - | - | yes |
| Caperna and Matthews 1996 | HIV | 588 | - | 63% | 91.8% | - | - | - | - |
| Essink-Bot et al. 1995 | Migraine | 846 | - | 63% | 90% | - | - | yes | yes |
| Humphreys et al. 1994 | Limb-threatening ischemia | 180 | 10 mins (by interview) | - | - | - | - | - | yes |
| Hurst et al. 1994 | Rheumatoid arthritis | 55 | - | - | - | - | - | - | yes |
| Hurst 1996 | Rheumatoid arthritis | 247 | - | 94.3% at baseline | - | yes | - | - | yes |
| Hollingworth et al. 1996 | knee problem | 102 | - | 89.2% at baseline | 83.3 at baseline | - | - | yes | yes |
| Norum and Angelsen[2] 1995 | Gastric cancer | 26 | - | - | - | - | - | - | - |
| Sculpher[3] 1993 | Menorrhagia | 200 | - | - | - | - | - | - | - |

1. All used except Brazier et al. 1993 2. Patients assigned to classification and scored by oncologists 3. VAS only

# Chapter 4

# The SF-36 Health Survey

This chapter begins with a detailed description of the features of the SF-36 health survey, its origins and why it is being considered in this thesis for use in economic evaluation. The SF-36 is then reviewed in the same way as the five preference-based measures were in the last chapter, namely in terms of the criteria of practicality, reliability, descriptive validity, validity of its values and empirical validity. The limitations of using it in economic evaluation are then examined. The next section appraises the alternative approaches to developing the SF-36 for use in economic evaluation and concludes with the approach selected for the research reported in this thesis.

## 4.1 The Short Form 36 health survey

### 4.1.1 Description

The SF-36 health survey is a standardised general measure of health status. It generates scores across eight dimensions of health: physical functioning, role limitations due to physical problems, role limitations due to emotional problems, social functioning, bodily pain, vitality, mental health and general health perception (Ware and Sherbourne, 1992). The dimensions of physical functioning, role limitations (due to physical problems), social functioning, pain, and mental health are most similar to the EQ-5D out of the preference-based measures.

The questionnaire comprises 36 items that are designed for self-completion and interviewer-administration by telephone and face-to-face. The content of the questionnaire has been summarised on Table 4.1 and reproduced in Appendix 2. There are between two and 10 items per dimension. Physical functioning has the most with ten, and bodily pain and social functioning the least with two. There are 35 items in all for the dimensions and a health transition item that is not included in the scoring. Respondents are asked to complete all items. For each item, the respondent has a choice of responses on a Likert scale. The items of physical functioning, for example, have

three response choices of: 'Yes limited a lot', 'Yes limited a little', and 'No not limited at all'. The items of mental health dimension have a choice of six responses from 'All the time' through to 'None of the time'.

In contrast to preference-based measures such as the EQ-5D, most of the items do not have any obvious ordinal relationship (e.g walking one hundred yards is not necessarily more important than climbing a flight of stairs). As a result, the 35 items of the SF-36 define a total of $2592 \times 10^{19}$ unique health states[1].

There are three components to the scoring of the SF-36 (see Appendix 3 for a detailed description of the scoring system). First, item responses are coded onto an equal interval scale increasing with health. For example, the three responses to the physical functioning dimension of 'limited a lot', 'limited a little', and 'not limited at all' are coded one, two, and three respectively. The six responses to the mental health items are coded from one for 'feeling calm and peaceful none of the time' to six for 'feeling calm and peaceful all of the time'. The only exception to equal interval scaling is the first General Health Perception item where the codes are excellent = 5, very good = 4.4, good = 3.4, fair = 2.0, and poor = 1.0. [2].

The second stage involves summing the coded items to generate the raw scores for each dimension. Thus for physical functioning:

Raw Physical functioning score = 3a + 3b + 3c + 3d + 3e + 3f +3g + 3h + 3i + 3j (1)

Where 3a to 3j are the coded responses to the individual items. This formulae therefore assigns an equal weight to each item. For physical functioning the lowest possible raw score is 10 and the highest is 30.

---

[1] Calculated as the product of the number of items in each dimension to the power of the number of levels of each dimension (e.g. PF=$10^3$ and RL (P) = $4^2$ together define 16,000 states).
[2] This was justified by the developers in terms of *its* improved correspondence with the original long form version (Ware et al, 1993).

The final stage is to transform this raw score onto a scale of zero to 100 using the following formula:

[(Raw score - lowest score) / (Highest score - lowest score)] * 100 (2)

This positive linear transformation results in a score range of 0 to 100 for each dimension, which was regarded as more convenient by the developer. It does not alter the measurement properties of the scale (e.g. it is not a ratio scale).

The SF-36 generates a profile of eight scores. The mean scores of a sample of the general population in Sheffield are presented on Table 4.2. Mean scores are also presented for the following groups: recent attendees at general practice ( i.e. in the last two weeks), a sample of people over 75, people found to have depressive symptoms, attendees at a chest clinic diagnosed with chronic obstructive pulmonary disease (COPD), and attendees at a rheumatology clinic with osteoarthritis. These profiles provide a descriptive picture of the comparative health of these different groups. It can be seen from these data that recent attendees at general practice had worse health than the general population. The profile of scores also indicates that patients with COPD were worse in terms of their physical health (e.g. physical functioning and role functioning due to physical problems) but they were similar to the general population in terms of their mental health. Those who have revealed depressive symptoms in another questionnaire were found to have lower mental health scores on the SF-36. These scores can also be used to monitor changes over time and hence measure the outcome of health care interventions.

### 4.1.2 History and development

The SF-36 health survey has evolved out of two major research programmes in the USA. The first was the Health Insurance Experiment (HIE) undertaken at the Rand Corporation to examine the health status of individuals who enrolled onto different schemes of organising the delivery and finance of health care (Brook et al., 1983). The original long form used in this study contained 108 items, covering a broad array of functional status and well-being concepts. The second programme was the Medical

Outcome Survey (MOS), which examined how different aspects of health care affect outcome (Tarlov et al., 1983).

Items for these long form versions were based on the results of a review of the literature in the 1970s. The usefulness of these full-length health batteries was seen to be limited by their size, particularly if they were administered alongside condition-specific measures. Short single-item scales, on the other hand, were not regarded as covering a sufficient range of health domains and were found to be unreliable and insensitive, particularly for small groups in trials (Ware et al., 1993). The developers therefore sought to achieve the compromise of *"... a standardised health status survey that is comprehensive, psychometrically sound, and brief"* (Ware and Sherbourne, 1992). The SF-36 was constructed to reflect the eight most important health concepts found in the MOS and other health surveys. It is intended to be generic in the sense of not being age, condition or treatment specific (Ware et al., 1993).

SF-36 health survey was published by Ware and Sherbourne in the USA in 1992. It was adapted for use in the UK by a team in Sheffield lead by the author (Brazier et al., 1992). Minor alterations were made to the wording of six items for use with UK populations (for example 'blue' was changed to 'low' and 'block' to '100 yards'). These revisions to the questionnaire have become incorporated into the official UK SF-36 (IQOLA, 1994), and will be the version used in this thesis. The SF-36 has now become one of the most widely used measures of general health in clinical trials in North America and Europe, and has been translated into over twenty languages (IQOLA, 1996).

### 4.2 Why consider using the SF-36 in economic evaluation?

Preference-based measures have been available for over two decades (e.g. Torrance et al., 1972), yet they are still not widely used. The applications of QALYs in the evaluation of health care interventions, for example, have been limited (Backhouse et al., 1992) and are certainly not sufficient to provide a complete and up-to-date assessment of the cost-effectiveness of health technologies (Drummond et al., 1993).

The SF-36, on the other hand, has recently become one of the most commonly used outcome measures in clinical trials in the UK and North America[3]. Most of these trials have been designed to address clinical rather than economic questions. Yet even amongst researchers who are seeking to address 'cost-effectiveness' questions, there has been a reluctance to use preference-based measures.

This reluctance is in part the consequence of continued unfamiliarity with preference measures. Drummond and Davies (1991) have suggested it is the reluctance that results from the perceived additional burden from using them in clinical trials, which in turn increases the costs of the trial and risks over-burdening the patient. In defence of preference-based measures, many of the instruments used by economists take less time than many clinical measures (e.g. the EQ-5D takes less than 3 minutes to complete), and cost considerably less.

A more fundamental concern with preference measures is their insensitivity and irrelevance for many conditions. There was some basis for this concern found in the review of the five measures in chapter 3. The interest in using the SF-36 in the research reported in this thesis comes from it being a rich and apparently sensitive measure of health across a range of domains.

Whatever the reasons for researchers not choosing to use one of the existing preference-based measures, the SF-36 has the potentially important advantage from being commonly used. The next question is whether the data it generates are likely to be useful in economic evaluation and this is addressed in the next two sections.

### 4.3 Review of the SF-36

This section reviews the SF-36 against the criteria used in the previous chapter to review the five preference-based measures of practicality, reliability, descriptive validity, validity of its values and empirical validity. It has not been based on a systematic review

---

[3]In the first year the UK Health Outcomes Clearing House, it had received more enquiries about the SF-36 than any other health measure (Outcomes Briefing, 1994).

owing to the large size of the published literature (a recently published bibliography identified xxx publications).

### 4.3.1 Practicality

The SF-36 is mainly used as a self-completed questionnaire, though there are interview and proxy based versions. With 36 items, it is longer than the Rosser, EQ-5D, HUI I-III, and 15D, but shorter than the QWB. The developers claim it takes 10 to 15 minutes (Ware et al., 1993), though it can take longer with elderly people (Hayes et al., 1995).

Response rates in postal surveys of general population samples in the UK have been between 72% (Jenkinson et al., 1993) and 83% (Brazier et al., 1992) and in a postal survey of patients with four common clinical conditions, exceeded 75% (Garratt et al., 1994). Among patient groups recruited in outpatients, rates have been over 90% (Brazier et al., 1996b; Harper et al., 1997). The rates of completion of all 36 items in these studies has exceeded 95%, though it has been found to decrease to between 68% and 89% by dimension in patients over 75 (Brazier et al., 1996a). Results are confirmed in applications in the USA (McHorney et al., 1994). For scoring the SF-36, it is not necessary to complete all items since the developers recommend a method for imputing responses by taking the average of the responses to the completed items within a dimension provided at least half of the items have been completed (Ware et al., 1993). In the group of elderly patients, this increased the proportion of usable responses (i.e. those that could be scored) to between 84.9% to 95.7%.

### 4.3.2 Reliability

In a survey of general practice patients, test and re-test (at two weeks) scores of the SF-36 dimensions were found to have rank correlations of between 0.60-0.81 (Brazier et al., 1992), which is within the range of 0.5 to 0.7 regarded as acceptable for group comparisons (McDowell and Newell, 1987). Furthermore, mean differences between test and re-test did not exceed one point on the 100 point scale, and plots of these differences against the subjects' scores did not reveal any bias over time. The correlations for COPD patients were also within the acceptable range for all dimensions except role limitations owing to emotional problems and social functioning. In a study

of patients with varicose veins, Garratt et al., (1994) report intra-class correlation coefficients between test and re-test at two weeks above 0.70 for all dimensions except Role limitations due to emotional problems.

### 4.3.3 Descriptive validity

Content and face validity

The content validity of an instrument depends on the relevance and appropriateness of its domains and items. The SF-36 is a measure of general health and not specific to any one medical condition. It is intended to measure the 'core' features of health (Ware et al., 1993). The dimensions were selected to cover the World Health Organisation's definition of health of a *"state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"* (WHO, 1948). The areas it does not cover in a direct way are specific symptoms (such as breathing or sleeping difficulties), and it excludes important health problems such as poor eye sight, or deafness. It covers 57.9% of all unprompted responses to the West Midlands survey of the general population conducted by Williams and colleagues at York (van Dalen et al., 1994). This compares to 35.9% for EQ-5D, 26.9% for the Rosser, 49.1% for SIP, 58.6% for NHP and 58.6 for QWB.

The items of the SF-36 were based on reviews of the literature and existing instruments at the time (i.e. 1970's), and refined by teams of academics over a long period of development. Items were selected to include positive as well as negative aspects of health, thus the mental health dimension includes 'Have you been a happy person?' as well as 'Have you felt downhearted and low?' (UK version) (MOS, 1993). In selecting items for the SF-36 from the long form versions, the developers employed a number of criteria (Ware et al., 1993). One criterion was to compare the performance of the shortened versions of the dimensions against the long form in detecting differences between populations. The SF-36 was therefore developed to correspond as closely as possible to the original scales in the long form and retain as much of the sensitivity as possible. The items have been selected and refined for use on adults. The items of the physical functioning dimension, for example, relate to common activities undertaken by

people in their daily lives. The main criticism of the face validity of the SF-36 has been the apparent lack of relevance of some of the items to very elderly populations (Hayes et al., 1994; Brazier et al., 1996a).

An attraction of using the SF-36, particularly on populations with mild health problems, has been the low proportion of respondents at the ceiling of the score range compared to other general measures of health, such as the Nottingham Health Profile and the EQ-5D (Brazier et al., 1992,1993). Conversely there have been concerns about possible floor effects in the physical dimension (Bindman et al., 1990).

Construct validity

The descriptions of the SF-36 have been validated for a wide range of common conditions by comparing SF-36 scores between groups. The dimension scores have been found to be able to show the expected patterns of scores between the general population and groups defined by age, recent use of health services, social class and chronic illness (Brazier et al., 1992), patients attending clinics with COPD and osteoarthritis (see Table 4.2), end stage renal disease (White et al., 1996), and the four common clinical conditions of varicose veins, peptic ulcer, back pain and menorraghia (Garratt et al., 1994). It has also been able to show significant differences within condition: by clinical severity in the COPD patients, and to distinguish transplant and dialysis patients.

The SF-36 has been shown to correlate significantly with other generic measures of health status, including the Nottingham Health Profile (Brazier et al., 1992), and the Sickness Impact Profile (Read et al., 1987; Katz et al., 1992); and with many condition-specific measures, including the Chronic Respiratory Questionnaire (Harper et al., 1997), the WOMAC osteoarthritis index (Brazier et al., 1996b), and the General Health Questionnaire-12 (McCabe et al., 1997), and the Arthritis Impact Measurement Scale (Katz et al., 1994). The SF-36 has also been reported to respond to health improvements found between preoperative and post-operative assessments in patients undergoing total hip arthroplasty (Katz et al., 1994). A UK study of patients comparing SF-36 with the EQ-5D and two condition-specific outcome measures for patients with Chronic Obstructive pulmonary disease (COPD) found most of the dimensions of the SF-36 to

be responsive to self-perceived health change in this patient group. The single index derived from the EQ-5D classification was found to be insensitive to these perceived changes.

The SF-36 has been compared to the Nottingham Health Profile (NHP), until recently one of the most commonly used profile measures of health status, and an earlier version of the Euroqol, the EQ-6D (Brazier et al., 1993). Considerable agreement was found between the SF-36 and the EQ-6D and the NHP, but the frequency distributions of the NHP scores and EQ-6D responses were significantly more skewed than for the SF-36 dimension scores. The skewness of the distribution of scores to the NHP reflected the fact that all item responses were dichotomous. Responses to the EQ-6D were skewed from the limitation of having only one item per dimension and two to three response categories for each item. As reported in Chapter 3, individuals who responded as having no health problem on dimensions of the NHP or EQ-6D were sub-divided into those who had at least the median SF-36 score (i.e. better health) and those who scored less than the median on comparable dimensions (i.e. worse health). For both the NHP and EQ-6D, patients in the poor health groups were found to have a higher mean age, and there was also a higher proportion of women and a higher proportion of patients not in full-time employment than the better group. The poor groups were also more likely to have consulted a GP recently, attended outpatients in the last three months, or been an inpatient in the last year. This evidence suggests the SF-36 is more sensitive than the NHP and the EQ-6D classification at detecting more mild perceived health problems (Brazier et al., 1993).

### 4.3.4 Valuation

For assessing the overall effectiveness of a treatment there are going to be situations where it will not be possible to compare the effectiveness of interventions using a profile measure such as the SF-36. An intervention could result in an improvement in physical functioning compared with another, for example, but an increase in pain. One solution to this problem is to combine the dimension scores or item responses into a single index using an assumed set of weights. This has been attempted by research team at Brunel University who aggregated the Nottingham Health Profile (NHP) into a single

index to estimate the QALYs gained from a heart transplant programme (O'Brien et al., 1987). Three methods of aggregation were utilised: (i) the proportion of affirmative responses to the 38 statements in the NHP; (ii) weighting the affirmative responses by weights estimated by the NHP developers, using Thurstone's method of paired comparisons (Hunt et al., 1986); and (iii) using unitary statement weights within dimensions and then weighting the dimensions by their proportion of the 38 statements. Such arbitrary weighting schemes could easily be applied to the SF-36, but they would not generate an index that could be legitimately used in an *economic* evaluation, because neither the dimension weights nor the dimension scores have been based on people's preferences.

There are also questions about the arbitrary nature of the assumptions underlying the coding of responses and the scoring to generate the dimension scores. There is no reason to suppose, for example, that a patient perceives the intervals of the responses to items of the physical functioning dimension of the SF-36 of 'not limited at all' and 'limited a little' to be equivalent to the interval between 'limited a little' and 'limited a lot'. To take another example, the intervals for the item on how much bodily pain a person has had in the last four weeks are 'none' to 'very mild', 'very mild' to 'mild', 'mild' to 'moderate', 'moderate' to 'severe', and 'severe' to 'very severe'. This would imply that a reduction in pain from 'mild' to 'very mild' would be equivalent to a reduction from 'severe' to 'moderate'. Evidence presented later in the thesis suggests that individuals are unable to perceive a significant difference between 'very mild' and 'mild', but perceive a large and significant difference between 'moderate' and 'severe' pain (Chapter 8). The summing of items makes equally untenable assumptions. In the physical functioning scale, the item 'limitations in climbing one flight of stairs' is assumed to be of equal importance to 'limitations in walking more than one mile'. For someone living in a bungalow, limitations in walking would probably be regarded as a far worse problem.

These assumptions imply that the SF-36 scores cannot, or at least are very unlikely to be a cardinal measure of people's preferences. There have been further doubts about the ordinal properties of the scores, particularly over small changes in dimension scores.

Williams (1989) has gone so far as to suggest that the use of arbitrary weights in some health measures is so serious a defect that it is doubtful *'whether the positive or negative changes in .... scores .... can be unambiguously rated as improvements or deteriorations in health state if properly valued'.*

### 4.3.5 Empirical validity

Despite these theoretical concerns there is evidence of a positive, if weak, association with people's preferences. Tsevat et al. (1991) in a study of patients infected with HIV examined the relationship between measures at two points in time. The preference-based measures were TTO, VAS and the QWB scale, and the SF-36. For TTO, the strongest correlates (0.51 to 0.59) were with measures of physical functioning (the SF-36 physical functioning score, SF-36 role limitation score, and SF-36 vitality). For the VAS rating measures, the strongest correlations with health status measures varied between 0.51 (SF-36 physical functioning) to 0.66 (SF-36 general health). The QWB was most strongly correlated with the SF-36 physical functioning (.51), and the SF-36 vitality (0.68). The authors report that the modest correlation found between preference-based measures in this particular study fits with similar findings from other studies of both HIV-infected and non-HIV infected patients. Bosch and Hunink (1996) looked at the relationship between the SF-36 and TTO and SG values in patients with intermittent claudication (mild peripheral arterial disease). The correlation coefficients from 0.16 (pain) to 0.46 (mental health), for the SG the corresponding correlations ranged from 0.10 (pain) to 0.34 (social functioning).

The scores for the physical, social general health perception dimensions were significantly associated with the stated preferences implied by the change in perceived health reported by COPD patients attending a clinic at six months apart. These reported changes were not found for the single index derived from the EQ-5D.

Many of the group comparisons reported under descriptive validity indicate that the scores usually confirm the pattern of hypothetical preference such as: the comparisons between the general population and patient groups diagnosed with COPD, osteoarthritis, end stage renal disease, peptic ulcers, back pain and menorraghia; the differences found

by clinical severity in COPD, and the differences between transplant and dialysis patients and the health improvements found between pre-operative and post-operative assessments in patients undergoing total hip arthroplasty.

4.2.6 Overview - key points

- It is brief and easy to use.
- There is evidence for its re-test reliability.
- Its dimensions cover the areas of physical, mental and social well-being, but exclude many symptoms and specific problems.
- There is some evidence of its ability to describe mild health problems and changes in health better than other generic instruments.
- The scoring of the SF-36 is not based on preferences, but arbitrary assumptions.
- There is evidence of a low to moderate association between SF-36 dimension scores and preference measures and indicators.

## 4.4 Using the SF-36 in economic evaluation.

Despite these criticisms of the valuation of the SF-36, economists may be asked to evaluate the relative efficiency of different interventions using results from the measure. This section examines the limitations to using the SF-36 in its current form in economic evaluation.

The usefulness of the SF-36 in assessing the relative efficiency of interventions depends on the results of an economic evaluation (Donaldson et al., 1996). Seven scenarios of costs and outcomes in a comparison of two interventions are presented in Table 4.3. The first scenario is a case of dominance where one treatment is cheaper and better on at least one of the dimensions of the HSM while being no worse on any other. In the second scenario the assessment of cost-effectiveness is also straightforward since it is simply a question of choosing the treatment with the better dimension scores since the two have been found to cost the same. The third scenario is the same across all dimensions of the SF-36 and hence it is a *cost-minimisation analysis* (CMA). Even for these three scenarios it is necessary to demonstrate the ordinality of the scale of the

dimension scores in relation to preferences. The theoretical reasons for doubting SF-36 dimension scores possess this property were reviewed in an earlier section. However, the empirical evidence shows they are significantly, if poorly correlated to preference measures. This suggests they should be able to rank states in the same order as preference-based measures, provided there is no trade-off to be made between dimensions. The SF-36 can be useful in assessing cost-effectiveness under each of these three scenarios.

The result is less straightforward for scenarios 4 to 7, where the usual technique for assessing relative efficiency would be *cost-effectiveness analysis* (CEA). The convention in CEA has been to measure health effects in natural units (Drummond et al., 1987) but Feeny et al. (1990) have suggested '*... the assessment of alternative drug regimens for the control of chronic respiratory disease could be displayed in terms of a set of cost-effectiveness ratios of the dollar per change in the CRQ score for each drug regimen.*' ( Where CRQ is the Chronic Respiratory Questionnaire, a condition-specific measure). However, Feeny and his colleagues point out that *'For specific and generic profile instruments that do not provide a single score, the meaningfulness of cost-effectiveness which utilises such measures is dubious'*. The problem arises from having multiple cost-effectiveness ratios. To assess cost-effectiveness it is necessary for a treatment to have a lower cost-effectiveness ratio across <u>all</u> dimensions of the HSM, otherwise it will be necessary to undertake trade-offs between dimension scores which are beyond the scope of these scales. In scenarios 4 and 5 one treatment performs better on the same dimensions but worse on others, and hence one treatment could be more of cost-effective on some dimensions but worse on others. Furthermore, our review of the evidence suggests that the dimension scores do not possess the interval properties required to generate cost-effectiveness ratios.

Where there are multiple outcomes, the recommended approach is to present the costs and benefits of the alternatives in a disaggregated form in a Cost Consequences Analysis and where there is no attempt to combine multiple outcomes into a single indicator of value. The decision-maker is left with the task of weighing up the costs against the multiple outcomes. This is a commonly used method of economic

evaluation. In the past this method has been known as a 'soft' cost-benefit analysis, but more recently has been called cost consequences analysis (CCA) (Drummond 1994). This approach can be seen as within the decision-aiding tradition of economic evaluation (Sugden and Williams, 1978). It has the advantage of retaining the way of thinking and discipline of economic evaluation.

CCA is likely to be particularly unhelpful with SF-36 dimension scores because they have no obvious intuitive meaning. The developers of the SF-36 acknowledge '*...when multiple items are combined into a score,...,the score has no inherent meaning.*' (Stewart and Ware, 1992). Score differences cannot be compared between dimensions, nor can dimension scores be compared to other outcomes (such as survival) or cost. The SF-36 may not be able to assist decision-makers in determining the relative cost-effectiveness of the interventions in such circumstances.

This section has described the limited circumstances where SF-36 may have a role in assessing relative efficiency. The usefulness of the SF-36 in economic evaluation depends on the results of the study. It is usually not possible to predict the results of a study and therefore the current advice to researchers designing an economic evaluation is to use preference-based measures alongside the health status measures such as the SF-36 (Brazier et al., 1997). However, the advantage the SF-36 has over the preference-based measures in terms of descriptive validity would be lost to the economic evaluation, and besides, this advice may not be taken up by other health services researchers. The alternative strategy is to adapt the SF-36 for use in economic evaluation.

### 4.5 Adapting the SF-36 for use in economic evaluation

After their attempt to adapt the NHP for use in economic evaluation using a number of arbitrary weighting schemes O'Brien et al. (1987) argued that '*a more formal process is required for translating health profile information, be it from the NHP or SIP with their richness and multi-dimensionality, into relative valuations of typical health states,*

*which can then be used to indicate relative quantity/quality of life trade-offs or preferences'.* This is the main purpose of the research reported in this thesis.

There are four possible approaches for incorporating preferences into the SF-36: 1) to map the SF-36 onto one of the five preference-based measures of health; 2) to estimate an exchange rate between dimensions of the SF-36 and one of the preference measures; 3) to use the descriptive data collected by the SF-36 to construct vignettes; or 4) to estimate values for a multi-dimensional scale constructed from the SF-36 These will now be reviewed to determine which would be the best approach for this research.

### 4.5.1 Approaches

Mapping

An important attempt at mapping health state descriptors from health status measures onto preference-based measures was undertaken by Gudex (1986). She mapped groups of scores from the Ruesch Social Disability Rating Scale (RSDRS) onto the Rosser matrix, thereby allowing outcome data for patients receiving maintenance haemodialysis to be converted into QALYs. For example, the 'social modifiers' dimension of the RSDRS was converted into the Rosser distress category using the following decision rule: a social modifiers score of 1-5 is equivalent to A on the Rosser distress scale, 6-19 is equivalent to B, 20-39 is equivalent to C, and 40-55 is equivalent to D. This rule is created solely by matching comparable descriptive states from the two scales and can therefore, *at most*, claim to possess face validity. This approach was used in several service settings with only limited success in producing cost per QALY data (Gudex, 1986). The mapping of patients onto the classification from other questionnaires has been found to be of questionable value since the process is necessarily based on a large number of arbitrary assumptions (Coast 1992; Drewett et al., 1992).

Exchange rates

This would entail estimating a relationship between dimension scores and a preference-based measure. Cairns and colleagues (1991) have explored the potential for establishing an exchange rate between the scores of three condition-specific measures to facilitate cross-programme comparisons: Montgomery and Asberg depression rating

scale, the Health Assessment scale for rheumatic disease, and Spitzer's Quality of Life Index for patients with Cancer and other chronic conditions. A sample of scenarios describing hypothetical patients was selected from these measures and a group of raters (n=66) was asked to rank and assign an index number to each using a VAS. The raters produced rankings for the health states by VAS that were in line with the original scales of the three instruments (Cairns and Johnston, 1992). However, the differences found in the VAS ratings were not constant between the intervals of these scales. This was evidence that there is no simple proportional relationship between the three condition-specific measures, or between these measures and preferences, at least as indicated by VAS. This implies the need for a large number of scenarios to be valued in order to estimate a non-linear functional form for the relationship or a complete revaluation of these instruments.

There is little other research on this topic, but the low to moderate correlations found in comparisons of SF-36 dimension scores with preference-based measures also suggest this is not likely to be a promising avenue for development. Furthermore, there is no theoretical reason for believing there will be a sufficiently consistent relationship between dimension scores and preference measures to have a valid and reliable exchange rate.

Valuation of Vignettes derived from the SF-36

This approach has the advantage of being able to focus on those aspects of health most relevant to the treatment being evaluated. Vignettes have usually been constructed from expert opinion or qualitative interviews with patients, rather than evidence collected in clinical trials using standardised questionnaires (e.g. Cook et al., 1994). The methodology for constructing the vignettes from the descriptive data generated by the SF-36 or any health status measure has not been developed.

A further drawback with this approach is that separate valuation studies would have to be undertaken for each economic evaluation. This has important cost implications in undertaking an evaluation. This might also reduce the usefulness of such studies for cross-programme comparisons since the respondents would be different. A further

potential cost and source of inflexibility would arise from sensitivity analyses, since the descriptive data are locked into the vignettes. To test other parameter values, such the range in the size of the improvements across dimensions, would require a re-valuation of the vignettes. This risks overloading the respondent.

Valuation of a multi-dimensional scale derived from the SF-36

Another solution is to estimate values for a multi-dimensional scale based on the SF-36 using preference elicitation techniques; in other words, to convert the SF-36 into a preference-based health measure.

Three 'measurement strategies' for undertaking this task have been identified by Froberg and Kane (1989b): holistic, whereby all health states defined by the classification are valued directly; explicitly decomposed; and statistically inferred decomposed. The second and third strategies require only a sample of health states to be valued. For the smaller instruments, such as the Rosser Disability/Distress matrix, direct valuation is feasible since it forms only 29 health states. The SF-36 would be too large for respondents to value all of its possible states. The explicit decomposition strategy was used to value the HUI The dimension scales are first valued on their own to derive single dimensional utility scales and then a sample of health states is valued to estimate the dimension weights by using multi-attribute theory. The statistical strategy has been employed to value the QWB and EQ-5D based on the use of regression to estimate weights from a sample of valued health states.

There are, however, potential problems with applying these decompositional solutions since the SF-36 is far larger and more complex than existing preference-based measures. The SF-36 was not designed with this task in mind. The developers of the SF-36 have admitted "*..the application of standard health state preference weighting procedures (e.g. Standard Gamble, Time Trade-off, multi-attribute theory) to obtain an overall score is not feasible*" (Hays et al., 1993). As shown earlier, item responses have no obvious ordinal relationship within dimensions. As a result, all the 35 items of the SF-36 would have to be included in the health states for valuation, and this would define

a total of 2592 x $10^{19}$ unique health states. The task of estimating a function for such a large and complex classification would be beyond the ability of the explicitly and statistical decomposed strategies. There would simply be too many states to value, and more importantly it would be impossible for respondents to value health states containing 35 items. Experience from other fields, such as transport, suggests people are only able to understand between 5 to 8 dimensions (Fourkes and Wardman, 1988). All except one of the existing preference-based measures have eight or less dimensions. These are major methodological problems to overcome.

### 4.5.2 Choice of approach

The mapping of the SF-36 onto a preference-based measure has been dismissed since it involves making arbitrary and in many cases dubious assumptions. The second approach has been rejected because there is no reason nor evidence to support a sufficiently consistent relationship between dimension scores and preferences to be able to estimate a valid and reliable exchange rate. The third approach of constructing vignettes from the SF-36 for valuation would overcome these concerns, but it would be expensive, potentially very demanding of patients, and the results would not be comparable between studies. The fourth approach of valuing a multi-dimensional scale based on the SF-36 has been chosen. Despite the methodological problems, it is preferred to the other three since it would, in principle, produce a preference-based health measure that could be used in multiple economic evaluations, and the results could be used to inform resource allocation decisions between programmes as well as within a patient group. This approach raises a number of major methodological problems, and these are addressed in the next chapter.

## 4.6 Conclusion

The Short Form-36 (SF-36) Health Survey is the subject of this thesis. It was therefore important to understand more about this measure, its characteristics and the reason for selecting it for this thesis.

The SF-36 is an important measure of general health and one of the most commonly used in clinical trials in the UK, the rest of Europe and North America. The self-administered instrument has been found to be practical in terms of its ease of use, high levels of response and completion and to be reliable at re-test. The strength of this measure lies in its descriptive validity, and in particular its sensitivity. Evidence was found of its greater sensitivity compared to the Rosser and the EQ-5D at detecting milder conditions and at responding to health changes in some groups of patients. The SF-36 is potentially a rich source of data for economic evaluation, but only has a limited use in assessing cost-effectiveness because the scores are not based on preferences. It is impossible to evaluate the relative cost-effectiveness of interventions when trade-offs must be made between dimensions of the SF-36 and/or cost. In these circumstances, the results could be incorporated into the framework of a cost-consequences analysis, but this would be of limited help to decision-makers given the difficulties in interpreting the scores.

The recommendation to use a preference-based measure alongside the SF-36 in clinical trials means the advantages of the SF-36 over the preference-based measures in terms of descriptive validity would be lost to the economic evaluation, and may not be taken up by many health services researchers. The adaptation of the SF-36 by incorporating preferences has the potential to extend considerably the application of cost-utility analysis in health care. This adaptation is the subject of the research reported in this thesis.

Four approaches to incorporating preferences into the SF-36 have been examined in this chapter: to map items of the SF-36 onto a preference-based measure; to estimate exchange rates for converting these scores into preference values; to construct vignettes from the results of each trial and to value these using one of the preference elicitation techniques; and finally to value a multi-dimensional scale based on the SF-36. The fourth approach has been chosen, but the valuation of such a large and complex descriptive classification as the SF-36 raises major methodological problems. These problems are addressed in the next chapter.

**Table 4.1 Dimensions, items, responses, and summary of content for the UK SF-36 health survey**

| *Dimensions* | *No. of items* | *Summary of Content* | *No. of response choices* | *Range of Response choice* |
|---|---|---|---|---|
| Physical Functioning (PF) | 10 | Extent to which health limits physical activities such as self-care, walking, climbing stairs, bending, lifting, and moderate and vigorous exercises | 3 | 'Yes limited a lot' to 'no, not limited at all' |
| Role Limitations - Physical (RP) | 4 | Extent to which physical health interferes with work or other daily activities, including accomplishing less than wanted, limitations in the kind of activities, or difficulty in performing activities | 2 | Yes/No |
| Bodily Pain (BP) | 2 | Intensity of pain and effect of pain on normal work, both inside and outside the home | 5 & 6 | 'None' to 'very severe' & 'not at all' to 'extremely' |
| General Health (GH) | 5 | Personal evaluation of health, including current health, health outlook, and resistance to illness | 5 | 'All of the time' to 'none of the time' |
| Vitality (VT) | 4 | Feeling energetic and full of life versus feeling tired and worn out | 6 | 'All of the time' to 'none of the time' |
| Social Functioning (SF) | 2 | Extent to which physical health or emotional problems interfere with normal social activities | 5 & 6 | 'Not at all' to 'extremely' & 'All of the time' to 'none of the time' |
| Role Limitations - Emotional (RE) | 3 | Extent to which emotional problems interfere with work or other daily activities, including decreased time spent on activities, accomplishing less and not working as carefully as usual | 2 | Yes/No |
| Mental Health (MH) | 5 | General mental health, including depression, anxiety, behavioural-emotional control, general positive affect | 5 & 6 | 'All of the time' to 'none of the time' |
| Reported Health Transition | 1 | Evaluation of current health compared to one year ago | 5 | 'Much better' to 'much worse' |

**Table 4.2: Mean SF-36 dimension scores**

| Patient group | PF | RP | BP | GH | VT | SF | RE | MH |
|---|---|---|---|---|---|---|---|---|
| General population[a] | 87 | 83 | 79 | 72 | 61 | 87 | 82 | 73 |
| Recent GP attendees[a] | 81 | 67 | 68 | 63 | 52 | 76 | 73 | 66 |
| People with depressive symptoms[b] | 72 | 52 | 59 | 53 | 39 | 62 | 41 | 53 |
| Elderly female (>75)[c] | 47 | 43 | 58 | 59 | 53 | 75 | 62 | 72 |
| COPD | 29 | 18 | 53 | 29 | 34 | 45 | 44 | 64 |
| OA of the knee (medical)[d] | 21 | 12 | 35 | 56 | 41 | 51 | 42 | 68 |
| Hernia repair[e] | 84 | 75 | 71 | 78 | 67 | 90 | 89 | 79 |

Sources:

a. Brazier et al,1992
b. McCabe et al, 1996
c. Brazier et al, 1996a
d. Brazier et al, 1996b
e. Lawrence et al, 1995

**Table 4.3: Assessing the relative cost-effectivess of two interventions given different cost and dimension scores on the SF-36**

| Scenario | Cost | Dimension scores | Can cost-effectiveness be evaluated? |
|---|---|---|---|
| 1 | Lower | Better in at least one dimension and no worse on any other | Yes, by dominance* |
| 2 | Same | Better in at least one dimension and no worse on any other | Yes* |
| 3 | Lower | Same across all dimensions | Yes, by cost-minimisation* |
| 4 | Lower | Better on some dimensions and worse on others | No |
| 5 | Same | Better on some dimensions and worse on others | No |
| 6 | Higher | Better in at least one dimension and no worse on any other | No |
| 7 | Higher | Better on some dimensions and worse on others | No |

* With the provisos about the ordinality of the scales

# Chapter 5

# Methods of Research

The valuation of a multi-dimensional classification scale based on the SF-36 using preference elicitation techniques involves three tasks. The first is to reduce and simplify the content of the SF-36 to form a multi-dimensional health state classification suitable for valuation. The second task is to value a sample of health states defined by the classification. The third task is to estimate values or weights for multi-dimensional classification from the sample of health states in order to be able to value all possible health states defined by the classification. This chapter describes the methodology employed to undertake each of these tasks, along with their rationale.

## 5.1 Adaptation of the SF-36

The ability of respondents to value reliably a set of health states reflects a combination of factors, including the size and complexity of the classification to be valued. Research has shown that individuals can only process between five and nine pieces of information at one time (Miller, 1956; Pearmain et al., 1991). This would imply a maximum of nine dimensions per health state, though given the complexity of some of the items of the SF-36, the number should probably be closer to five. It would also imply each dimension should only be represented by one simple statement in each health state. This implies a multi-dimensional classification with ordinal dimensions, where one item is selected from each dimension to define a health state. This is the structure of classifications of the Rosser scale, HUI, EQ-5D and 15D. The size and structure of SF-36 must be radically altered to meet this requirement.

The desired size of the classification is further constrained by the ability of respondents to discriminate between the levels of a dimension. In transport economics, such qualitative variables have usually been restricted to two or three levels (Bradley, 1988), though there are examples of five in market research (McCulloch and Best, 1979). The desire to keep the number of levels per dimension to a minimum, however, must be

weighed against the need to retain the sensitivity of the original instrument. It is the greater sensitivity of SF-36 over existing QALY instruments that is a major reason for this research.

The aim is to produce a classification based on patient responses to the SF-36 health survey. The new classification will not be an additional questionnaire. It must be possible, therefore, to map SF-36 responses onto the new classification in order to apply it to existing data sets. It is also important for the text of the SF-36 items to be altered as little as possible.

The construction of a multi-dimensional health status classification from the SF-36 involves many judgements. It is important to be explicit about the basis for these judgements.

### 5.1.1 The process

The adaptation of the SF-36 was undertaken by a research team based in Sheffield consisting of a sociologist, a general practitioner, a research psychologist, a statistician and the author acting as co-ordinator (Brazier et al., 1994). A series of meetings was held by the team in October 1993 - February 1994. Between the meetings, individual team members were assigned a dimension and asked to consider the options for adapting into an ordinal structure and prepare supporting evidence. At the meetings, the team collectively arrived at decisions based on the judgement criteria set out below.

### 5.1.2 Judgement Criteria

The aim was to produce a multi-dimensional health state classification from the SF-36 with five to six ordinal dimensions. To this end, the team used the following guidelines in making its decisions:

1. Redundancy was to be avoided.

Where two or more items appeared to be describing the same aspect of health and were found to be closely correlated, only one item would be selected. Similarly, response

choices were merged if there was evidence from the IQOLA survey that respondents regarding them as equal (see below).

2. Exclusion of positive items.

The SF-36 has items designed to measure positive as well as negative aspects of health (e.g. "did you feel full of life?" "did you feel worn out?" respectively). The team decided to exclude the positive items since they were judged to be less relevant in informing resource allocation in health service provision.

3. People's preferences

There was not the time nor the resources to conduct an extensive survey of people's attitudes to each item of the SF-36, and therefore this was often not known. However, there was an opportunity to use the results of a survey undertaken for the IQOLA project into the relative value of different statements in the questionnaire using VAS (Ware et al., 1995)[1]. The purpose of the IQOLA survey was to test the translation of the instrument from American-English to UK-English, and not to inform this exercise (for which it was incomplete). Nonetheless, it did provide some useful evidence on preferences that has been used to determine whether to retain certain items and whether to merge response choices.

5.1.3 The decisions

<u>General Health Perception</u>

An overall assessment of health is an important dimension in a profile instrument such as the SF-36. Where the aim is to generate a single health index it would be illogical to include a general health dimension. It was therefore decided to exclude this dimension from the revised classification.

---

[1] The sample was 102 individuals selected on a convenience basis in Sheffield and included health professionals, students and members of the general public (Brazier et al, 1994). They were asked to indicate the relative importance of the categories of each of the six response choices of the SF-36 on a VAS anchored by the two most extreme choices (e.g. 'none of the time' and 'all of the time'). There were an additional nine questions asking respondents to indicate the relative importance of a selection of items on a VAS from within a dimension (e.g. limited in bathing and dressing and limited in bending, kneeling or stooping).

Physical function

There are 10 items in this dimension, each with three response choice categories (i.e. 3a to 3j on the SF-36, see Appendix 4.1). These items can be divided into four sub-dimensions: activities (3a and 3b), climbing stairs (3d and 3e), walking (3g, 3h and 3i) and bending (3f and 3j). The item 'lifting or carrying groceries' (3c) cannot be classified into any single one of these sub-dimensions. The three response categories for each item make it impossible to rank the items *a priori*, even within these sub-dimensions. To simplify the problem, it was decided to merge the response 'Yes, limited a lot' with 'Yes, limited a little' into a single category of 'limited'. This may reduce the sensitivity of the scale, and this is examined later in the thesis (Chapter 9). The revised cna now be ranked within the four sub-dimensions as follows:

| | |
|---|---|
| Activities: | No limitation |
| | Limited in vigorous activities |
| | Limited in moderate activities |
| Climbing: | No limitation climbing several flights of stairs |
| | Limitation climbing several flights of stairs |
| | Limitation climbing one flight of stairs |
| Bending: | No limitation bending, kneeling or stooping |
| | Limited bending, kneeling or stooping |
| | Limited in bathing and dressing |
| Walking: | No limitation in walking more than a mile |
| | Limited in walking a mile |
| | Limited in walking half a mile |
| | Limited in walking 100 yards |

The problem was how to form a single ordinal physical functioning scale from these sub-dimensions.

Not being 'limited in vigorous activities' was found to be the least severe limitation in the IQOLA survey. This item was therefore used as an end point of the scale, namely level 1 (Figure 5.1). Level 2 was therefore being 'limited in vigorous activity'.

In the IQOLA survey 'limited in climbing several flights of stairs' and 'limited in walking more than a mile' were given similar scores (i.e. 4.77 and 5.04 respectively on a VAS from 0 to 10) and so they have been combined to form level 3. The item 'being limited in climbing one flight of stairs' was combined with 'being limited in walking half a mile' to form level 4, though the equivalence of these two items was not tested in the IQOLA survey. These items have been combined in order to reduce the size of the scale.

Finally, being 'limited in walking 100 yards' was not directly compared with 'limited in bathing and dressing' in the IQOLA survey, and there are no *a priori* grounds for ranking one over the other. They have been ordered as level 5 and 6 respectively in the classification. This ordering may prove to be wrong for most respondents. This is not a problem provided most respondents agree on the ordering. In the presentation of health states in the valuation survey, the respondent is not constrained to this ordering of items.

The new scale for physical function excluded three items: 'moderate activities', 'lifting or carrying groceries', and 'bending, kneeling or stooping'. Responses to these items were found to be highly correlated with the original physical functioning scale (r = 0.62, 0.63 and 0.69 respectively after correction for overlap, Brazier et al., 1992) and the consequent information lost by excluding these items was considered to be small.

Role limitations

There are role limitations due to physical health and those due to emotional problems. In the original SF-36 these are separate dimensions, with four and three items respectively, and dichotomous response choices (yes or no). There is no *a priori* basis for ranking these items. In the IQOLA survey, the role (physical) items were ranked in order of severity, but the intervals were not even. Various ways of combining these items were considered, but none was found to be satisfactory. There is further complexity from having two dimensions. The developers of the SF-36 argued that role limitations due to physical problems could be distinguished as a separate concept from those due to emotional problems. The survey results supported this claim with comparable limitations due to emotional problems being valued more highly than those

due to physical problems (e.g. 7.5 and 5.6 for being unable to work or perform any other activity).

The two role dimensions of the SF-36 have been found to be the least satisfactory in terms of completion, particularly for the elderly (Brazier et al., 1992; 1996a). Respondents who do not work often believe these items are not relevant to them, and hence miss them out. On grounds of parsimony, the team decided it was not worth developing an elaborate scale for these dimensions, and therefore combined them into a simple two-level scale: "You have a problem/you have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems".

Social functioning

There are two items to this dimension. Different ways of combining these items were considered, but none was found to be satisfactory. It was therefore decided to select one, and the team opted for the item which asks 'to what extent...' rather than 'how much time...', since the former item is likely to be more general. All five response choices were retained as levels in the new scale since the mean ratings of the choices were equally spaced along the VAS in the IQOLA survey.

Pain

This dimension also has only two items, one with five and the other six levels. Again, there was no simple way of combining the items and one item had to be selected. It was decided to exclude the item referring to the interface with 'normal work'. Although it is explained in the SF-36 questionnaire that 'normal work' includes housework, there is inevitably scope for misunderstanding amongst individuals who do not work. The loss of information should be minimal given the large correlation between the two items (r = 0.74, Brazier et al., 1992).

The size of the intervals between the mean ratings of each of the six response choice categories in the IQOLA survey was sufficient to justify retaining them all.

Mental Health

There are five mental health items in the SF-36, each with six response choice categories (i.e. 'all of the time' to 'none of the time'). These items cover four concepts of mental health: anxiety (2), depression (2), behavioural control (1) and positive mental health (1) (Stewart and Ware, 1992). The positive items (9d and 9h) have been excluded from our version of the scale on the basis of the second judgement criterion. The behavioural/emotional control item was difficult to distinguish from depression, with a high correlation (0.66), and was therefore excluded on grounds of redundancy. The anxiety item has the lowest correlation with the other items, and along with depression is a major problem in mental health.

In common with the developers of the Health Status Index and Euroqol, the team decided to combine the depression and anxiety items into a single statement. This resulted in the statement 'You feel down hearted and low or you have been a very nervous person most of the time'. This combination was thought to be inconsistent and unnecessarily complicated. Previous work by McHorney and Ware (1993) had found that 'feeling tense' was an alternative anxiety item which could be used interchangeably for 'a very nervous person' in a long-term follow-up of patients. The original item of the SF-36 statement was therefore replaced by the alternative form. The new scale therefore reads 'You feel tense or downhearted and low ...'.

The intervals between the mean ratings of the response categories in the IQOLA survey did not suggest any obvious scope for merging categories. To reduce the size of the scale, however, it was decided to merge 'none of the time' with 'a little of the time' to form the upper end of this scale. The latter was thought to lie at the positive end of the mental health difficulties and therefore did not justify its own category.

Vitality

There are four vitality items (9a, 9e, 9g and 9i), each with the same six response choice categories as mental health. These items divide into those relating to distress or negative aspects of vitality (items 9g and 9i) and those concerned with positive health

(items 9a and 9e). In line with the selection criterion, only the former have been used. There was no basis for choosing between the distress items and therefore it was decided to combine them to form 'you feel worn out or tired... of the time'. The response choice categories were reduced to five on the same grounds as mental health.

### 5.1.4 The new multi-dimensional health state classification: SF-6D

These changes have created six ordinal dimensions with two to six levels (Figure 5.1). Together these form the Short-Form Six-Dimensional Health State Classification (SF-6D). A health state is composed of six statements, one from each dimension, starting with physical functioning and ending with vitality. There are a total of 9000 health states defined in this way (see examples in Figure 5.2). All responders to the SF-36 questionnaire can be assigned to the SF-6D provided the 14 items used in the six dimensions of the classification have been completed.

The following SF-36 data sets have been assigned to the SF-6D: the Sheffield general population sample, people found to have depressive symptoms, and attendees at a chest clinic diagnosed with chronic obstructive airways disease (COPD). (The SF-36 scores for these data sets were presented in Chapter 4 on Table 4.2). The results are shown as frequency distributions by dimension on Figure 5.3. The frequency distributions of the general population data are skewed towards the left-hand side (less severe end), with more than 50% of people in the least severe category on three of the dimensions. The frequency distributions for patients with COPD are very different. The functioning scales (physical, role and social) showed the greatest impairment compared with the general population results, while for the mental health and vitality dimensions the differences were less marked. In the sample with symptoms of depression, patients are shown to differ most in terms of mental health. These results provide some initial support for the descriptive validity of the new scales. The general population data have a more skewed distribution than the original SF-36 dimensions. The implications of this for the sensitivity of the measure and a comparison with EQ is presented in Chapter 9.

This new classification is the result of a trade-off between two competing considerations: the cognitive abilities of respondents and the desire to retain the descriptive richness of the SF-36. The success of the SF-6D in meeting these twin objectives will be examined in the empirical work presented in Chapters 6, 8 and 9. At this stage it is important to acknowledge that the SF-6D may not prove to be the best compromise between these objectives. The results of the valuation survey may suggest further revisions to the classification in terms of its content, number of levels, and number of dimensions. This will be addressed later in the thesis.

## 5.2 Valuation Survey

The following design issues must be addressed for the valuation survey: the preference elicitation technique to use; the version of the chosen preference elicitation technique to use; the selection of respondents to undertake the valuation tasks, the sampling of health states (out of the 9000 possible states) and the methods of data collection. These are now described.

### 5.2.1 The choice of technique for eliciting preference

The most commonly used techniques for eliciting preferences for health states are the visual analogue or rating scale (VAS), magnitude estimation (ME), time trade-off (TTO), equivalent numbers or person trade-off (PTO) and standard gamble (SG) (Torrance, 1986). There are supporters of each of the preference elicitation techniques: VAS by Kaplan and colleagues (Kaplan and Ernst, 1983); ME has been used to value the Rosser classification (Rosser and Kind, 1978); TTO is favoured by Richardson (1995), Johannesson et al. (1996) and Dolan et al. (1996); SG by Feeny and Torrance (1989) and Gafni and Birch (1993); whilst Nord (1992) has advocated the PTO method because it incorporates social values. The choice of elicitation technique is important because it has been shown that they generate different values (Bombardier et al., 1982; Dolan and Sutton 1997; Loomes et al., 1994).

Comparisons of these techniques have tended to focus on their basis in economic theory (Brooks, 1995). Theoretical basis is important to those economists who subscribe to the

conventional welfarist perspective. They will be concerned with a link between the preference elicitation techniques and consumer theory. This link seems to be less important for those who have rejected welfarism and favour a broader-based 'extra-welfarist' perspective (e.g. Richardson, 1994). Whichever the perspective, it is also important to consider the empirical properties of the measures including: practicality, data quality (e.g. the consistency of respondents' valuations), reliability, and if possible, empirical validity (Froberg and Kane, 1989b; Richardson 1994; Dolan et al., 1996). These properties have been described at length in the review of preference-based measures (Chapter 3.2).

This section has two parts. In the first part each technique is described, along with its theoretical basis. The second part compares the techniques in terms of their theoretical basis, practicality, consistency, reliability, and validity and two techniques are selected for this survey.

Visual Analogue (or rating) scale

Definition:

*"A typical rating scale consists of a line on a page with clearly defined endpoints. The most preferred health state is placed at one end of the line and the least preferred at the other end. The remaining health states are placed on the line between these two, in order of their preference, and such that the intervals or spacing between the placements correspond to the difference in preference as perceived by the subject"* (Torrance, 1986, P.18).

The distances between health states on a visual analogue scale should reflect a person's feelings about the relative differences in preferences between them on an interval scale. The differences in a person's feelings between 90 and 95 on the scale should be the same as between 20 and 25 (hence it is often called a 'feelings' thermometer).

To use it in calculating QALYs it is necessary to value death and perfect health alongside health states of interest. The raw ratings of each respondent are transformed using the following formula:

$$A_i = \frac{R_i - R(\text{death})}{R(\text{best}) - R(\text{death})}$$

where $A_i$ = Adjusted VAS rating for health state $h_i$
R (death) = Raw rating given to unconsciousness, followed by death
$R_i$ = Raw rating given to $h_i$
R (best) = Raw rating given to the best health state

This transformation results in the value 1.0 for the best health state and zero for death. The value of $A_i$ would lie within this range, or assume a negative value for states valued as worse than death. This adjustment is claimed to allow inter-personal comparisons (Torrance, 1986; MVH, 1994).

There has been an interest in the literature in using VAS as a technique for deriving a value function for preferences under certainty (Keeney and Raiffa, 1976; Dyer and Sarin, 1982; Broomes, 1993). Such a value function can, according to Dyer and Sarin (1979), be given a "strength of preference" interpretation, and they have established a formal theoretical basis for the link between a value function and utility theory (Dyer and Sarin,1982). Dyer and Sarin argued that an individual's utility function can be seen as a combination of a measurable value function and relative risk attitude. Specifically, every utility value is related to each value of the value function uniquely by a transformation for relative risk. They suggest the use of VAS as a means of measuring this strength of preference concept. This has important practical implications since VAS is a far easier technique to administer than some of the other techniques, such as SG and time trade-off, and this has led to an interest in estimating the relationship between these valuation techniques by estimating a power term for relative risk (Torrance, 1976; Loomes et al., 1994; Dolan and Sutton, 1997).

The theoretical basis of VAS as a method for eliciting preferences has, however, been disputed by many economists (Nord, 1992, Loomes et al., 1994; Richardson, 1994). The main criticism has been the absence of choice in the elicitation task. The rating scale does not confront the respondent with the notion of opportunity cost and hence there can be little confidence in the results indicating the economists' notion of strength

of preference. Interviews with respondents indicate this is not the meaning they intend (Nord, 1991a; Morris and Durand, 1989; Loomes et al., 1994). When asked respondents talk about concepts of fitness, or notions of chronology, not the value of health.

Magnitude estimation

Definition:

*"Here the subjects were asked to provide the ratio of undesirability of pairs of health states - for example, is one state two times worse, three times worse etc. compared to the other state? Then, if state A is judged to be x times worse than state A, the undesirability (disutility) of state B is x times as great as that of state A. By asking a series of questions all states can be related to each other on the undesirability scale"* (Torrance, 1986, P.25).

The ME is intended to generate a ratio scale. However, the extent to which it provides a ratio scale measure of strength of preferences has been criticised since, like the VAS technique, it does not confront the respondent with the opportunity cost of his/her choice. It measures feeling and attitudes, rather than strength of preference (Nord, 1992; Richardson, 1994). There is also no attempt to incorporate the consequences of risk (Loomes and McKenzie, 1989).

Nord (1992) has pointed out that the version of ME used by Rosser and Kind (1978) asked respondents to value death as well and hence they were explicitly considering the quantity/quality trade-off. In these lengthy interviews respondents were also invited to reflect on the implications of their ME valuations in terms of the proportion of resources to be allocated for the relief of one condition compared with another. Respondents were allowed to alter their valuations after this reflection[2]. The high values in the original Rosser matrix compared to VAS, better reflected the value given to life or the 'rule of rescue' (Nord, 1991a).

---

[2] The Rosser matrix was recently revalued by researchers at York, but they did not include this reflective stage in their interviews and obtained significantly lower values for each health state (Gudex et al, 1993). Surprisingly the York research team did not consider Nord's explanation (which at least provides a reason for the direction of the difference).

Standard Gamble

Definition:

*"The subject is offered two alternatives. Alternative 1 is a treatment with two possible outcomes: either the patient is returned to normal health and lives for an additional t years (probability P), or the patient dies immediately (probability 1-P). Alternatively 2 has the certain outcome of chronic state i for life (t years). Probability P is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is simply P; that is hi = P* (Torrance, 1986, P.20).

The SG technique is directly derived from expected utility theory (EUT)[3]. By setting death to zero and full health to one, 'P' is a cardinal index measuring an individual's utility under uncertainty (Von Neumann and Morgenstern, 1944). This index is unique up to a positive linear transformation and therefore an interval scale. The axiomatic basis of SG in the classical theory of decision-making under uncertainty seems highly relevant to medical decisions, and this has led to it being regarded by many health economists as a 'gold standard' amongst valuation techniques in health care (Gafni and Birch, 1993; Drummond et al., 1987; Torrance, 1986). Furthermore, this index incorporates the person's relative attitude to risk (Dyer and Sarin, 1982).

The alleged theoretical advantages of SG have been questioned by a number of economists (Richardson, 1994; Broome, 1993; Buckingham, 1993). These stem from evidence that the axioms of EUT have been shown to be violated in numerous studies (see brief review in Chapter 2.3.2). Richardson (1994) and Broome (1993) are particularly concerned about the influence of factors other than the preferences for a health state and relative risk, such as the gambling effect. As Richardson (1994) points out *"Von Neumann and Morgenstern did not believe that their axioms accounted for the specific utility from risk"* (P. 10). The SG procedure may introduce a specific utility from taking risk, but this is not allowed for in a N-M function. *"At worst it introduces*

---

[3] For states regarded by the respondent as worse than death, the gamble is changed. The outcomes in alternative 1 become either: the patient is returned to normal health or the patient lives in the state worse than death for an additional t years. Under alternative 2 the patient dies immediately.

*an additional random element whose relationship to the specific utility of the risk associated with a medical procedure is unknown"* (Richardson; 1994, P. 11).

Time trade-off

Definition:

*"The subject is offered two alternatives - alternative 1: state i for time t (life expectancy of an individual with the chronic condition) followed by death; and alternative 2: healthy for time x<t followed by death. Time x is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is given by hi = x/t''* (Torrance, 1986, P.23)[4].

TTO was developed by Torrance et al. (1972) as an alternative to SG for use in health state valuation. They regarded TTO as an approximation to the 'gold standard' of the SG technique and advocated its use as a proxy measure since it was thought to be simpler to administer. Torrance (1976) found TTO yielded the same values as SG (Torrance, 1976), but more recent studies have found significant differences (e.g. Dolan and Sutton, 1997). TTO has since been taken up by other economists, including Richardson (1994), Buckingham (1993) and the MVH group at York.

The TTO question asks individuals to make a choice and consider the opportunity cost of health status in terms of life years lost. Buckingham (1993) has suggested that TTO elicits values for a utility function with health status and length of life, among other things, in its arguments (see equation 7 in Chapter 2). Richardson (1994) has argued in favour of TTO since it provides a direct method of yielding a QALY (in contrast to SG), and is more meaningful to decision-makers.

The theoretical basis of TTO has been criticised for failing to take account of uncertainty in medical care. Patients are never likely to have a choice between x years

---

[4] TTO has been adapted for valuing health states regarded as worse than death. Here alternative 1 involves dying immediately. Alternative 2 involves x years in the health states regarded as worse than death followed by (t-x) years in perfect health. Again, duration x is varied until the respondent is indifferent between the two alternatives. The formula for calculating the health state value becomes - X / (t-x).

in full health for certain or y years in another state for certain. In reality, there is always uncertainty. Mehrez and Gafni (1991) therefore argue TTO does not generate a cardinal utility function at all, but a value function[5].

Person trade-off

Definition:

*"If there are x people in adverse health situation A and y people in adverse health situation B, and if you can only help (cure) one group (for example, due to limited time or limited resources), which group would you choose to help?". One of the numbers x or y can then be varied until the subject finds the two groups equivalent in terms of needing or deserving help. If x and y are the equivalent numbers as judged by the subject, the undesirability (desirability) of condition B is x/y times as great as that of condition A. By asking a series of such questions all conditions can be related to each other on the undesirability scale"* (Torrance, 1986, P. 25).

The PTO differs from the other techniques described in that it incorporates the notion of choice in a context of limited resources and social choice rather than the conventional individual choice. It has been argued by Nord et al. (1993) that this makes it more relevant for informing resource allocation decisions across programmes. The respondent is being asked to consider the value of a health state for another person, and this may be valued differently from a health state for themselves. PTO has resulted in higher values for states than either SG or TTO, and Nord (1992) has suggested this is because people are less willing to sacrifice the lives of others. As Richardson (1994) notes, the choice of PTO implies a judgement in favour of incorporating paternalistic rather than libertarian values. For Nord (1992) this is a virtue, since the task conforms more closely to the way the results will be used and therefore the 'meaning' or 'intent' behind the

---

[5] A further theoretical concern arises from the likely influence of time since the number of years (x) in health state hi exceeds the number of equivalent years in full health (t). For an individual with a positive time preference, the value of hi is reduced since (at least for states preferred to death) the number of years in this state (t) exceeds the number of years in full health (x). Assuming a constant positive discount rate, the application of discounting reduces x by a greater proportion than t, and hence the ratio of x to t (i.e. hi) is reduced. It is possible to adjust TTO for a constant time preference rate, but this is strictly outside of the QALY model.

respondents' valuations is more relevant. It is also likely to include judgements about equity, which again Nord regards as an advantage.

This technique has an intuitive appeal, though its theoretical basis has not been rigorously examined. There are important questions about the assumptions underlying this method and the implications for using the results outside of the context of the question. There is also no consideration of how time preference or risk should be incorporated. The appropriateness of using 'paternalistic' values may also be questioned.

## A comparison of techniques

### Theoretical basis

There is no basis in economic theory for the use of VAS or ME for deriving QALYs, since they do not present a choice and hence are not able to measure strength of preference on a cardinal scale. These techniques may generate ordinal information on preferences but the scores will not have the interval properties required for economic evaluation. There has been interest in estimating the relationship between VAS and SG or TTO, but this has had only limited success, particularly at the individual level (e.g. Torrance, 1976, Read et al., 1984, Loomes, 1993; Dolan, 1995a).

PTO, TTO and SG all confront the respondents with hypothetical choices, but PTO is different in that it presents social choices. This implies a judgement in favour of incorporating paternalistic values into the valuation of health dimensions. In Chapter 2, the case was made for limiting benefits to health, but as Culyer (1989a) has argued, this is not a reason to reject consumer theory. This is ultimately a value judgement.

The status of SG comes from its foundation in EUT, and many economists regard it as the only valid method for obtaining utility values. TTO is only regarded by its original developer as a proxy for SG, and recent evidence suggests it is a poor proxy (Feeny and Torrance, 1989). However, Richardson has argued that this view of SG as the only valid measure of utility depends on just one of at least four different definitions of utility

in economics: utility as a measurable psychological concept of welfare or well-being, utility as an ordinal set of preferences, utility as an index of strength of choice, and Von Neumann-Morgernstern's utility as an index derived from the SG. Primacy should not be given, therefore, to a measurement tool because of its apparent conformity to one particular theory of utility. Richardson recommends a broader set of 'extra -welfare' criteria for judging alternative valuation techniques based on their measurement properties (e.g. monotonicity and interval scaling) and their clarity of meaning to decision-makers. On these grounds he argued TTO (and PTO) provides more direct and more meaningful measures for decision-makers.

There is considerable disagreement in the health economics literature on the choice between SG and TTO for use in economic evaluation. An important dimension in this debate, however, concerns what is being valued. Some users of TTO have been interested in its use to value whole scenarios rather than health states (Richardson, 1994, Cook et al.. 1994 and Sculpher 1996). These scenarios can include uncertainties, such as risk of death and other adverse outcomes, as well as time. TTO may not incorporate risk in the valuation procedure, but it can reflect people preferences for the uncertainty contained in scenarios. The case for preferring SG would seem less compelling for valuing health scenarios. In the context of the QALY model, risk attitude towards health status is incorporated through the elicitation of utility values under conditions of uncertainty, namely SG (equation 8, Chapter 2). Furthermore, an attraction of SG is that it mirrors elements of medical decisions. On these grounds SG is preferred to TTO for deriving QALYs.

Empirical

This part of the review draws on the work of Froberg and Kane (1989b), who reviewed the methods of VAS, ME, SG, TTO and PTO against the criteria of reliability, validity and feasibility. Their review has been updated from two recent comparative studies undertaken at York; one a comparison of VAS, ME and TTO (Gudex et al., 1993) and the other comparison of SG and TTO (Dolan et al., 1996).

*Practicality*

Despite the theoretical concern with VAS, this technique has been widely used in health economics, including in the valuation of the QWB scale (Kaplan, 1989), and Euroqol (Euroqol, 1990), and this in part is due to the simplicity and brevity of the task (Froberg and Kane, 1989b). The acceptability of the technique to patients is reflected in its high rates of completion (MVH, 1994). The ME method involves a slightly more complex task, but there is no evidence of it being less acceptable.

The TTO method is thought to be simpler to administer than SG, since it does not involve the concept of probability, but the more familiar idea of time (Torrance, 1976; Froberg and Kane, 1989b; Nord, 1992)[6]. However, to make the notion of probability easier for respondents to understand, visual aids or props have been developed, such as the probability wheel (Torrance et al., 1992). Furthermore, rather than asking respondents to pick a point of indifference in one go, a 'ping-pong' technique is used, where the researcher asks the respondent to start at the extreme probability levels of 1.0 or 0.0. The researcher progressively narrows the probability range until the respondent arrives at a point of indifference. Another version of SG starts the respondent at the top end of the probability range, and asks respondents to indicate whether or not they would accept the gamble. The respondents then works their way down until they are not sure. The same process is undertaken from the other end. Similar methods have been developed to help respondents perform a TTO task.

In the York comparison, TTO achieved higher levels of completion than SG, with and without props (Gudex et al., 1993). However, the completion rates were very high for both techniques: TTO with props achieved 99.2%, TTO without props 95.8%, SG no props 95.6%, (using the second method) and SG with props 94.7%.(using the ping-pong

---

[6] I would argue that the SG task offers a more credible choice scenario than TTO, since the latter presumes there is always certainty SG can be presented in terms of a decision about whether or not to have surgery:
*"Choice A is a surgical procedure and choice B is to stay in the health state shown in the bottom box. Choice B is certain, but choice A (the surgery) is risky. It doesn't always work. If it does work, you will be in the health state shown in the left hand box in Choice A. But if the treatment does not work, you will die immediately, shown in the right hand box in Choice A."*
Adapted from Thomas and Thomson (1992).

variant). The version of SG without props was self-completed and hence it reduces the cost of data collection compared to interviewer-administered methods.

*Consistency*

1) Internal consistency

VAS has been found to produce fewer illogical rankings of health states than ME, TTO or SG (Gudex et al., 1993; Dolan et al., 1996). ME produced the most inconsistencies. In Nord's work, PTO was used to value two EQ-6D health states; and the logical ordering of the states was found to be inconsistent with the scale. He argued this was owing to the complexity of the states defined by the EQ-6D, yet this is a comparatively simple classification. This evidence suggests respondents had difficulty with the PTO task. TTO was found to result in fewer inconsistencies than SG with the EQ-5D scale, but this was not statistically significant (Dolan et al. 1996).

2) Consistency between versions

All valuation techniques generate values which differ according to the version used (Nord, 1992). For VAS, there is evidence of 'distribution' and 'contextual' effect. The distribution effect is the tendency for respondents to use all response categories equally (Stevens and Galanter, 1957), and result in 'spreading out' of responses. Contextual effects are where the average ratings for items are influenced by the level of other items being valued. The rating of the seriousness of offences by respondents has been found, for example, to be influenced by the relative seriousness of the other offences being valued (Parducci, 1983). Loomes et al. (1994) found a similar effect in health, whereby core health states had lower (and hence worse) values in the 'nice' group of states compared with a 'nastier' group. ME values have been found to be influenced by whether full health or some other state was used as the anchor point (Nord, 1992).

SG valuations are susceptible to changes in the reference points used in the gamble, for example replacing a fatal outcome with poor non-fatal outcome (Llewellyn-Thomas et al., 1982). Whether or not SG and TTO have been administered with props has been found to effect results significantly (Dolan and Sutton, 1997). Care must be taken in how all of these techniques are used.

*Reliability*

a) Inter-rater

In the review by Froberg and Kane (1989b) the levels of inter-rater reliability were found to be similar between VAS and ME, but markedly worse for PTO. One of the York studies found the ME method to be more vulnerable than VAS or TTO to an interviewer effect (Gudex et al., 1993).

b) Re-test

Froberg and Kane (1989b) found little evidence on re-test reliability. In the only study they found, re-test reliability at one week or less was found to be similar between SG, TTO and VAS. The York comparison of SG and TTO resulted in the highest correlations between test and re-test being achieved by TTO with props. The differences were significant in comparison to SG with props (0.83 vs. 0.54) and to TTO without props (0.83 vs. .90) but not compared to SG without (0.83 vs. 0.74). Performance is more dependent on the version used than the technique itself.

*Empirical Validity*

There is evidence of convergent validity, whereby the values generated by the different techniques are significantly correlated, but there are substantial differences between the techniques (Torrance, 1976; Bombadier et al., 1982; Llewellyn-Thomas et al., 1984; Read et al., 1984; Bass et al., 1994). However, this evidence cannot be used to discriminate between techniques.

Selection of technique

The theoretical advantages of SG in deriving health state utilities for the QALY model must be weighed against the evidence of the York study on completeness, consistency and reliability. The version of TTO with props performed significantly better than the SG with props. However, the differences were not significant in the comparison with SG non-props. Furthermore, what little evidence is available from other studies, suggests TTO does not perform better against these criteria than SG. It has therefore been decided to use SG in the valuation survey.

The only argument for using VAS is to provide proxy values for one of the choice-based methods. Statistical models relating VAS to SG and TTO have had some success in relating mean health state values (Torrance, 1976). VAS has been shown to explain a substantial portion of variance in SG and TTO, but no single functional form has been dominant. This could be because there is no single theory to explain the relationship.

There are important reasons for having a second elicitation technique in the survey. It reduces the dependence on a single technique. There are only sufficient resources to undertake one survey, and should problems arise with the SG, then it seems a safe policy to have a second. It is also common practice to precede valuation procedures such as SG (or TTO) with a simpler valuation task in order to familiarise the respondent with the health classification and the concept of value (e.g. Jones-Lee et al., 1993; MVH, 1994). The addition of TTO or PTO would overburden the respondent and hence the choice is between VAS and ME. VAS would seem to dominate in terms of completeness, consistency and reliability, and is easier to administer than ME.

### 5.2.2 Choice of versions of SG and VAS

There remains the important choice of the version of each technique to use. The method of administration was constrained by the resources available for this survey. There were also insufficient resources for piloting the different versions. Therefore, it is important to select questionnaires which have already been shown to be acceptable, consistent and reliable, and require few resources to administer.

#### Visual analogue scale

There are numerous versions of the VAS with different designs and methods of completion (McDowell and Newell, 1987). The most widely used in health economics has been the thermometer design adopted by the Euroqol group, and recently used in the MVH survey of over 3000 homes (Euroqol, 1990; MVH, 1994). For this version respondents are given a set of health states in no particular order. They are asked to place the states of health along a vertical 10cm line with endpoints of 'best imaginable' and 'worst imaginable' (see Figure 5.4). The intervals between the states on the rating

scale should reflect the differences between the states. To assist them in completing the task, respondents are asked to begin by ranking the health states.

This version of the VAS has a number of important methodological advantages. It can be self-completed with little or no explanation. There are fixed and well-defined end points on the thermometer i.e. 'best imaginable' and 'worst imaginable'. These should, according to Kaplan and Ernst (1993), minimise the risk of a spreading effect. A context effect can be avoided by ensuring respondents receive a balanced combination of good and bad health states. It has been extensively tested in York, and found to achieve excellent levels of completeness, consistency and reliability. An additional advantage arises from being able to compare the results of this survey with those obtained in the MVH survey of the UK.

Standard Gamble

The variant developed by Torrance (1976; 1986) involved the aid of a probability wheel. The subject is asked to iterate between extreme values for the probability of success P, such as 100% and 1%, towards a point of indifference (i.e. the 'ping-pong' method). An alternative variant was recently developed by Jones-Lee and colleagues (1993) without the use of a visual aid. Instead it uses a questionnaire with a list of values for chances of success. From this list, subjects are asked to indicate all the values of P where they are confident they would choose the treatment and all the values where they would reject treatment. Finally, they are asked to indicate the value where they find it hardest to decide.

Torrance (1986) argued that the use of props reduced inconsistency. However, the York MVH team (Dolan et al., 1996) recently conducted a pilot study comparing a sliding scale prop (instead of a probability wheel) utilising the 'pin-pong' method, and a self-completed questionnaire based on a version developed by Jones-Lee et al. (1993). Completion rates for the two versions were very similar at 5.3% and 4.4% respectively, but consistency was slightly better for the non-props version, though the difference was not statistically significant, and there were no differences in terms of re-test reliability. The self-completed version used in the MVH pilot survey has therefore been selected,

since it is less costly to administer. An additional advantage of this version is that it asks subjects to give upper and lower bounds to their answers. Jones-Lee and colleagues argue that most people are likely to experience difficulty in giving precise answers to questions about points of indifference.

The probability categories listed in the MVH version of the questionnaire were all five points apart on the zero to 100 scale. For the very mild states, this limits people to choosing either five in 100 risk of death or no risk at all. This may have been less of a problem for the EQ-5D classification since it defines more severe states, but for many states of the SF-6D it is unlikely that people would not agree to a 5% risk and therefore the scale would be insensitive to mild states. To improve the sensitivity at the upper end of the scale, additional categories of chances of success were introduced of 0.96, 0.97, 0.98 and 0.99. The number of categories at the other end of the scale was correspondingly reduced. As in the original Jones-Lee version, respondents also had the opportunity to indicate whether they would only choose the treatment if it had a lower than one in 100 chance of failure and to indicate at which level they would accept the treatment.

The SG question used in the survey is shown on Figure 5.5a and b. In the instructions, respondents were asked to imagine that they were in a chronic state of ill-health. A successful outcome of treatment results in a better state of health, but failure will lead to unconsciousness followed shortly by death. They were asked to consider a range of chances of success starting from 100 in 100 down to 10 in 100 (with a final box for immediate death preferred). Respondents were first asked to indicate with a tick all those chances of success where they would choose the risky treatment. Then they were asked to place a cross against cases where they would reject the treatment, starting from 10 in 100. Finally they were asked to indicate where it was most difficult to choose.

For ethical reasons it was decided in the patient questionnaire to change the worst treatment outcome in the SG question from death to a non-fatal health state for most of the gambles. Many of the patients in the sample had terminal conditions, and the health professionals responsible for their care felt it was inappropriate to confront them with

death in every question. According to EUT, it should be possible to translate the health state values onto a scale of death to perfect health, providing the treatment failure health state has itself been valued on this scale. Two of the eight gambles undertaken by patients were therefore the conventional gambles involving death. This also provided an opportunity to test the predictions of EUT that a directly valued health state, where the treatment failure is death, should be equivalent to one estimated indirectly from two gambles.

In the MVH study, a version of this questionnaire was used for health states regarded as worse than death, and the choice was between death now or a treatment which might result in full health or the state worse than death. It was thought this would have made the booklets too large for this survey. Respondents were therefore asked to consider whether treatment failure would be preferred, but no probability value was obtained. SF-6D health states are less severe than those defined by EQ, and this omission proved to be of little consequence: only 16 out of 1747 ratings were negative on the VAS, and none for SG.

In the VAS and SG questions, respondents were told the chronic health states would last for ten years. Under the QALY model, the duration specified in the question should not influence the value given to a health state, owing to the assumption of a constant proportional trade-off between life-years and health. Ten years was used in the MVH survey.

### 5.2.3 Respondents

The question of whose preferences or values should be used in valuation surveys is a value judgement. Respondents who have experienced the health states could be argued to be in a better position to understand the states (Buckingham, 1993) and are likely to be the most immediate recipients. Doctors and other health professionals might be thought to have a broader view, since they see a range of conditions. On the other hand, perhaps it should be a representative sample of tax-payers or the electorate. There are arguments for any of these constituencies, and they have all been used in QALY research (Torrance, 1986).

The aim for this research was to select a group of respondents which reflected the different groups commonly used in valuation surveys. Care was taken to obtain representatives of the general public, patient groups, health care professionals and health care managers/administrators. Given the limited resources available to undertake this study, it was not possible to use a systematic method of sampling. Instead, convenience samples were drawn from students on two health economics courses (including NHS managers, clinicians and nurses, and undergraduates), medical school staff, and patients attending hospital outpatient clinics in respiratory medicine, rheumatology and a centre for diabetes care.

Sample size was determined by available resources, rather than any formal calculations. In the original study proposal, funds had been obtained for 120 respondents, but in the event this was extended to 165. This number compares favourably with the sample sizes of many past studies in this field [e.g. Rosser and Kind, 1978 (n=77), Grogono and Woodgate, 1971 (n=25); Torrance, 1982 (n=112); Sintonen, 1981, (n=77)], but it would be regarded as small in comparison with large scale national surveys (e.g. MVH, 1994).

### 5.2.4 Health States

The 110 non-patient and 55 patient respondents in this survey are able to undertake 1760 valuations (i.e. 110 non-patients x 12 plus 55 patients x 8). It would have been possible for all respondents to have valued different states, but this would have raised concerns about the reliability of the value for each state. It would also not have been possible to examine the effect of respondent background characteristics and taste. A compromise was reached between the desire for a reasonable spread of states and the need for reliable estimates, so that in the event 58 SF-6D health states were chosen and perfect health.

The selection of health states should be undertaken by factorial designs, where the choices are determined by the statistical desire to avoid multi-collinearity. Applications in transport economics have shown that this often results in unrealistic combinations of attributes (Fowkes and Wardman, 1988). Similar problems may arise for the SF-6D; for

example, very severe pain being combined with no problems on other functioning dimensions such as role limitations and social functioning. These types of combination will lack credibility with respondents. Fowkes and Wardman (1988) suggest that any cost in terms of multi-collinearity will be outweighed by the enhanced realism.

Nonetheless it is important to achieve a good balance of SF-6D dimensions and levels in the sample of health states used in the valuation survey. All levels of each dimension must appear, and preferably more than once, in the selection of health states. To be able to examine interactions in the modelling, it is also important to have different combinations of levels of the six dimensions. Health states were therefore chosen to include those with predominantly physical problems (i.e. physical functioning, role limitations and social functioning) or mental health problems (e.g. mental health and energy) and combinations of physical and mental health problems. States were also selected for being 'mild' or 'severe'. However, it was also important for states to be plausible to the respondents. To ensure all health states were plausible, it was decided to limit the selection to those states which occurred in surveys using the SF-36 questionnaire data sets for the following patient groups: chronic obstructive airways disease, hernia, cholecystisis, menorrhagia and mental illness and a general practice sample (Brazier et al., 1994). These were selected because they were the ones available at the time.

The 58 health states (including full health i.e. health state 111111) are presented on Table 5.1, and the resultant distribution across levels on Table 5.2. The distribution is not equal between levels, but most levels appeared five or more times. The notable exceptions were level 5 of social, level 6 of pain and level 5 of mental health which occur 4, 2 and 3 times respectively, reflecting the infrequency with which such extreme health problems occur, even in populations with conditions sufficiently serious to be referred to hospital.

Experience from past studies suggests the maximum number of states or alternatives that can be valued at any single administration by a member of the general public is between 9-16 (Kroes and Sheldon, 1988; Dolan, 1995). In this survey, it was decided a

non-patient respondent would value only 12 and patients to value just eight health states.

All respondents were asked to undertake three valuation tasks. The first was to rank the health states and the second was to rate them on a VAS. The patients ranked and rated a single sample of eight health states together. For the non-patients, the 12 states were divided into two groups for valuation. The third task was to value each health state using SG.

To meet the competing objectives of survey, some states were valued by more respondents than others. Three were valued by all or most of the respondents, 16 'common' states valued by approximately one in four, and the remaining 50 states valued by one in 10 (Table 5.1). Health states were divided into blocks of five health states. The blocks with 'common' states were A to D and with 'rare' states were blocks E to N. Both sets of blocks contained one of the core states, and the remaining four were either 'common' or 'rare' states respectively. All blocks contained a balance of 'good' and 'bad' states. Each respondent was allocated one of the common blocks and one of the rare blocks.

At the respondent level it was important to ensure each individual faced a balanced set of health states. The review in Chapter 3 found evidence of VAS ratings for health states being influenced by the presence of other states in the valuation exercise (i.e. the 'range frequency model' of Parducci, 1983). The balance between 'good' and 'bad' states in each block of states should help avoid this problem.

### 5.2.5 Methods of data collection

Each respondent was given a booklet containing questions on personal background, self-rated health and the valuation exercises (see Appendix 5.1 for an example of a booklet). The background questions covered age, employment, occupation and industry, age on completing full-time education and health. Health was assessed by the first item of the SF-36 (i.e. would you say your health is excellent, very good,.., poor), a question from the General Household Survey on long-standing limiting illness, and the SF-6D itself.

Asking respondents to state their own health in terms of the six dimensions of the SF-6D was also a useful way of introducing them to the scale. These questions were followed by one ranking and rating exercise for patients, and two for non-patients. The SG questions came last. Ranking and rating were presented first in order to familiarise the respondents with the health states and the notion of valuation before moving onto the more complex SG task. There is also evidence of an 'ordering effect' and therefore it was important for all respondents to undertake the tasks in the same sequence (Llewellyn-Thomas et al., 1984).

The majority of respondents completed the questionnaire in groups of three to 15 in supervised sessions. At the beginning of a session, respondents had the purpose of the survey explained to them by a researcher, and were carefully taken through examples of each exercise. These explanations were read by the researcher from a prepared set of notes (see Appendix 5.2). Respondents were encouraged to seek clarification both before and during the valuation exercises. Only two researchers were involved in the supervision of these sessions (i.e. Rosemary Harper and myself).

## 5.3 Modelling health state values for the SF-6D

There are two strategies for extrapolating values for the whole multi-dimensional classification system from a sample: the explicitly decomposed method based on algebraic solution using multi-attribute utility theory (MAUT) and the other is decomposition by statistical inference (Froberg and Kane, 1989a; Currim and Sarin, 1984). These two strategies have been examined in the context of the comparison of the preference-based measures, and in particular the HUI and EQ-5D (Chapter 3). Essentially, MAUT enables the values of all health states defined by a multi-dimensional classification to be calculated from: 1) the valuation of single dimensional scales and 2) the valuation of a set of multi-dimensional 'corner states'. Assumptions about the functional form of the relationship between the dimensions of health. This is usually additive or multiplicative function (Foot note 5, chapter 3), since in practice it is

usually not possible to estimate the more general multi-linear form. The dimension weights are calculated by solving a set of equations.

The statistical strategy involves the estimation of a model of the relationship between the dimension levels of the classification and value or utility scores using regression-based techniques. The simplest functional form is the additive, with each dimension level being entered as a dummy variable. More complex forms would incorporate interactions between the levels of the dimensions. The size of the model will be limited by the degrees of freedom available, but the model-building techniques available in econometrics enable the analyst to ensure there is no redundancy in the model in the form of insignificant interactions. This has been the approach adopted in transport economics (Bates, 1988). It is able to incorporate error into the model, test the model's assumptions, and utilise a range of well-established techniques in econometrics.

The algebraic model is deterministic, and offers no means of testing the fit of the model. A rigorous testing would require a substantial number of additional health states to be valued, since this would allow comparison of actual with predicted. This may be the reason for the statistical inference substantially outperforming the algebraic method in its ability to predict the choices of respondents over different multi-dimensional states. Currim and Sarin (1984) found the correlation between actual and predicted choices over jobs with different mixes of attributes was 0.16 for the algebraic methods and 0.64 by statistical inference from using von-Neumann and Morgenstern utility values and assuming an additive function. This has been confirmed in comparisons of predicted health state values with the values elicited from the same respondents for the HUIs and 15D (Torrance et al. 1992; Sintonen 1994b). For these reasons, statistical inference is the chosen strategy for estimating health state values for the SF-6D.

The detailed methodological issues raised by this strategy are addressed in Chapter 8.

## 5.4 Conclusion

The adaptation of the SF-36 for valuation has required a trade-off between the cognitive abilities of the respondent and the desire to retain the sensitivity of the original instrument. The result is the SF-6D. The results of the valuation survey may suggest further improvements.

Standard gamble (SG) has been chosen as the main technique for eliciting preferences, and VAS the secondary technique. The self-completed versions developed by the MVH group from earlier work have been used, although the SG question was adapted for the milder states defined by the SF-6D compared to the EQ-5D. This has the advantage of allowing some comparison of results with that survey. A convenience sample of 165 respondents has undertaken the tasks on a selection of 58 health states. Statistical inference has been chosen to estimate values for the SF-6D.

The results from this survey and the subsequent statistical analyses are presented in the next three chapters.

**Figure 5.1: SF-6D**

Physical functioning

1. Your health does not limit you in vigorous activities (e.g. running, lifting heavy objects, participating in strenuous sports).
2. Your health limits you in vigorous activities (e.g. running lifting heavy objects, participating in strenuous sports).
3. Your health limits you in climbing several flights of stairs or in walking more than a mile.
4. Your health limits you in climbing one flight of stairs or in walking half a mile.
5. Your health limits you in walking 100 yards.
6. Your health limits you in bathing and dressing yourself.

Role limitation

1. You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems.
2. You have problems with your work or other regular daily activities as a result of your physical health or any emotional problems.

Social functioning

1. Your physical health or emotional problems do not interfere at all with your normal social activities.
2. Your physical health or emotional problems interfere slightly with your normal social activities.
3. Your physical health or emotional problems interfere moderately with your normal social activities.
4. Your physical health or emotional problems interfere extremely quite a bit with your normal social activities.
5. Your physical health or emotional problems interfere extremely with your with your normal social activities

Bodily pain

1. You have no bodily pain.
2. You have very mild bodily pain.
3. You have mild bodily pain.
4. You have moderate bodily pain.
5. You have severe bodily pain.
6. You have very severe bodily pain.

Mental health

1. You feel tense or downhearted and low a little or none of the time.
2. You feel tense or downhearted and low some of the time.
3. You feel tense or downhearted and low a good bit of the time.
4. You feel tense or downhearted and low most of the time.
5. You feel tense or downhearted and low all of the time.

Vitality

1. You feel worn out or tired a little or none of the time.
2. You feel worn out or tired some of the time.
3. You feel worn out or tired a good bit of the time.
4. You feel worn out or tired most of the time.
5. You feel worn out or tired all of the time.

**Figure 5.2: Health states defined by the SF-6D**

Dc

Your health limits you in climbing several flights of stairs or in walking more than a mile.

You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems.

Your physical health or emotional problems interfere moderately with your normal social activities.

You have mild bodily pain.

You feel tense or downhearted and low a good bit of the time.

You feel worn out or tired a good bit of the time.

Db

Your health limits you in climbing one flight of stairs or walking half a mile.

You have problems with your work or other regular daily activities as a result of your physical health or any emotional problems.

Your physical health or emotional problems interfere moderately with your normal social activities.

You have no bodily pain.

You feel tense or downhearted and low some of the time.

You feel worn out or tired some of the time.

Da

Your health limits you in bathing and dressing yourself.

You have problems with your work or other regular daily activities as a result of your physical health or any emotional problems.

Your physical health or emotional problems interfere moderately with your normal social activities.

You have moderate bodily pain.

You feel tense or downhearted and low some of the time.

You feel worn out or tired most of the time.

De

Your health limits you in walking 100 yards.

You have problems with your work or other regular daily activities as a result of your physical health or any emotional problems.

Your physical health or emotional problems do not interfere at all with your normal social activities.

You have moderate bodily pain.

You feel tense or downhearted and low a little or none of the time.

You feel worn out or tired some of the time.

Dd

Your health does not limit you in vigorous activities (e.g. running, lifting heavy objects, participating in strenuous sports).

You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems.

Your physical health or emotional problems do not interfere at all with your normal social activities.

You have mild bodily pain.

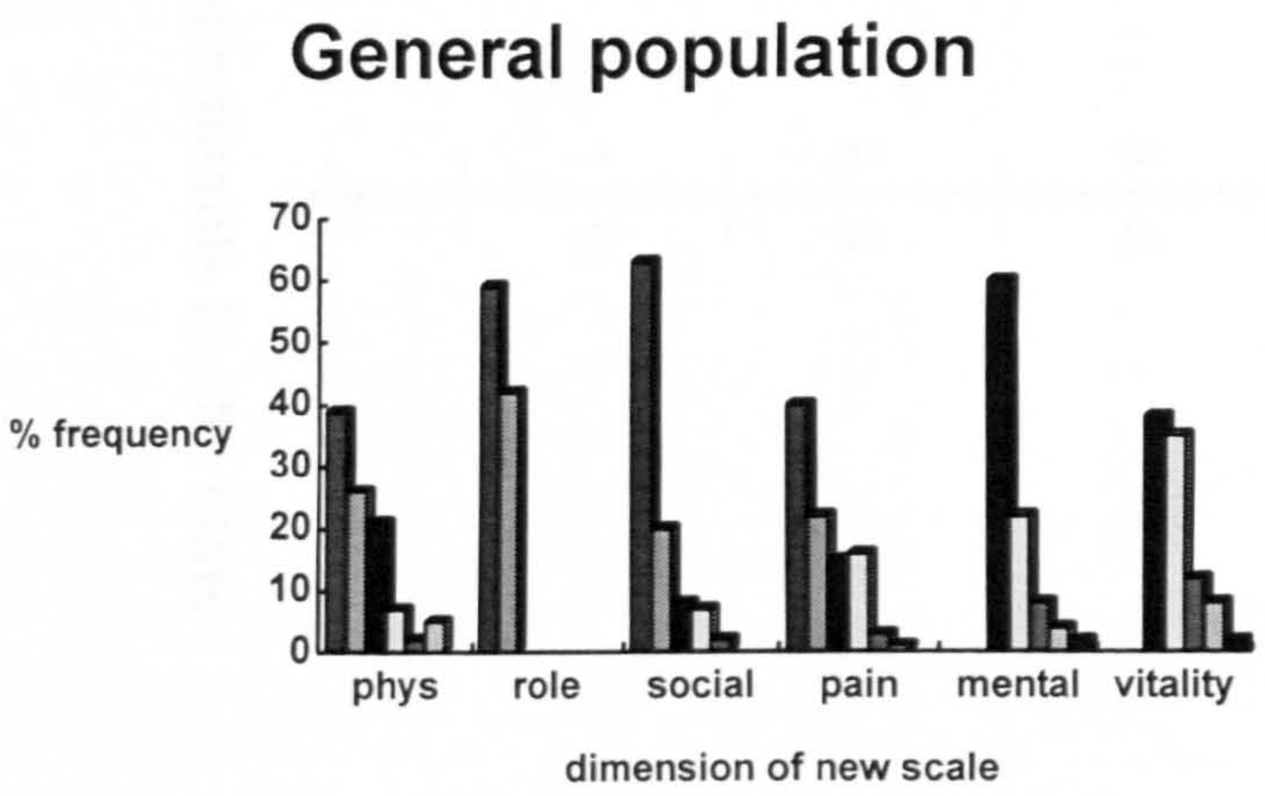
You feel tense or downhearted and low a little or none of the time.

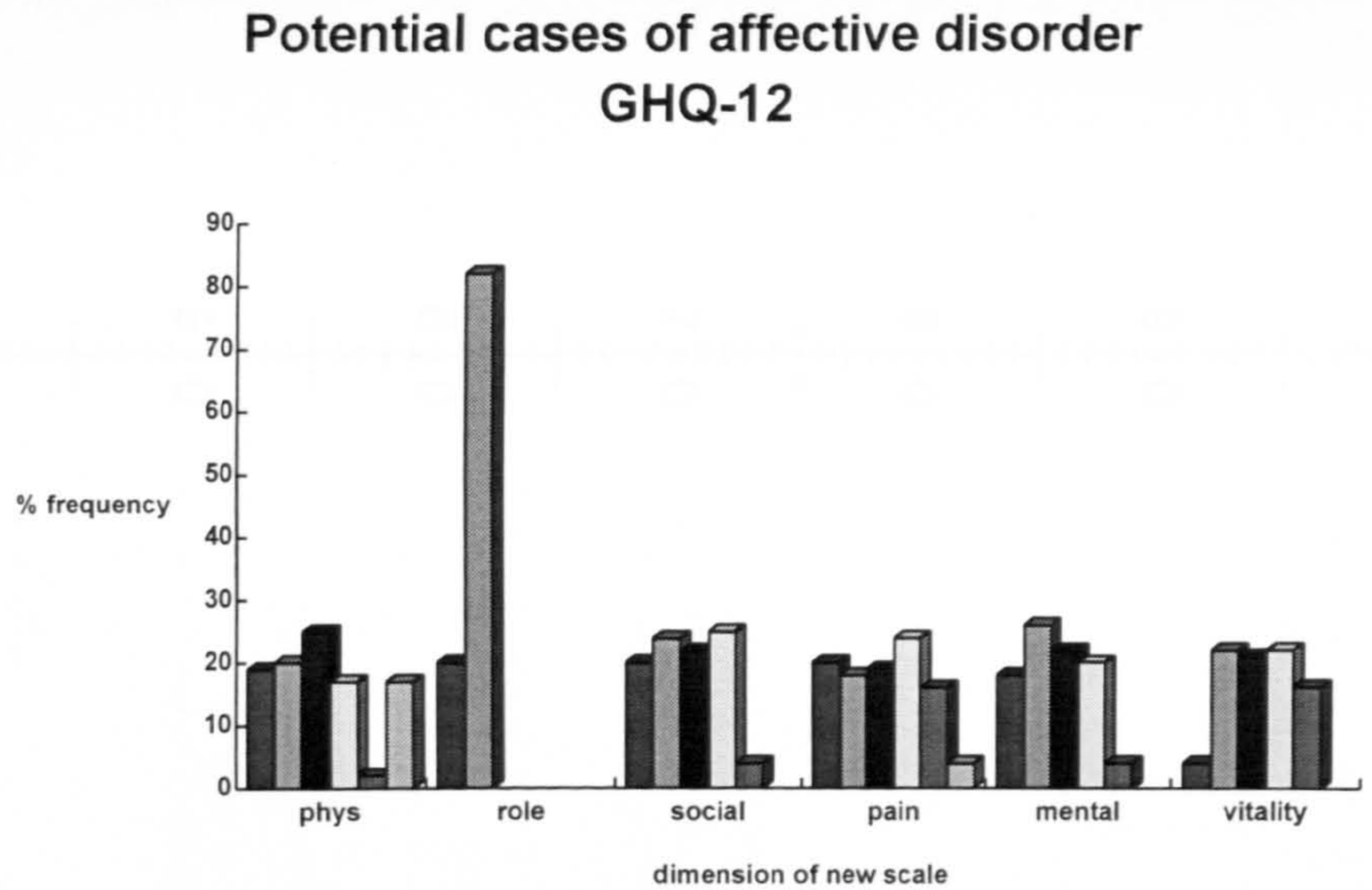
You feel worn out or tired some of the time.

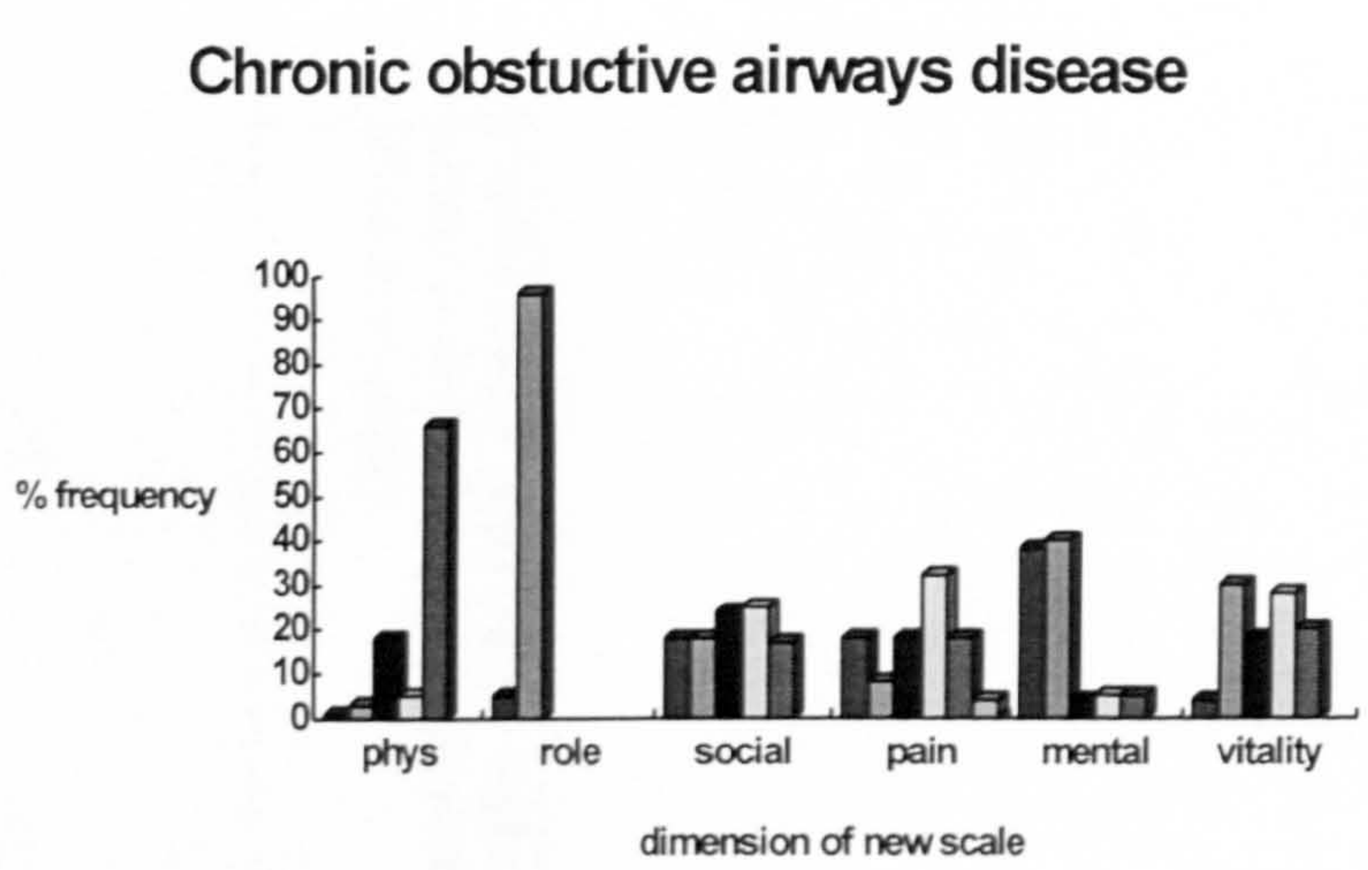
O

Unconsciousness followed shortly by death.

**Figure 5.3: Frequency distributions of the general population and patient groups across dimension**



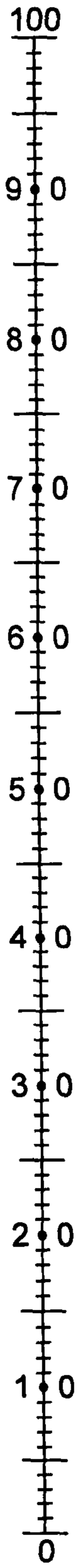




**N.B. The dimension categories are left to right: best through to worst**

**Figure 5.4: The visual analogue scale or ('thermometer') question used in the valuation survey**

**Best imaginable health state**




**Worst imaginable health state**

**Rating exercise 1**

Now please indicate the relative positions of each of the health states on this scale.

(It may be helpful to mark your ratings of the best and the worst health states first, followed by the intermediate states)

Please repeat the ranking and rating exercise with the cards in Envelope II
(See over page)

**Figure 5.5a: A standard gamble question used in the valuation survey.**

Suppose you were in a state of ill-health shown immediately below. The doctor tells you that you will remain in this condition for TEN years unless you have treatment. However, this treatment does not have a certain outcome. If it succeeds, it will result in a better state of health. If it fails, you will shortly die. The choice is therefore between:

**FOR CERTAIN**

**Ba**

Your health limits you in bathing and dressing yourself.

You have problems with your work or other regular daily activities as a result of your physical health or any emotional problems.

Your physical health or emotional problems interfere moderately with your normal social activities.

You have moderate bodily pain

You feel tense or downhearted and low some of the time

You feel worn out or tired most of the time.

**OR**

**IF TREATMENT SUCCEEDS**

**P**

Your health does not limit you in vigorous activities (e.g. running, lifting heavy objects, participating in strenuous sports).

You have no problems with your work or other regular activities as a result of your physical health or any emotional problems.

Your physical health or emotional problems do not interfere at all with your normal social activities.

You have no bodily pain.

You feel tense or downhearted and low a little or none of the time.

You feel worn out or tired a little or non of the time.

**IF TREATMENT FAILS**

**O**

Unconsciousness followed shortly by death.

**Figure 5.5b: The standard gamble answer sheet used in the survey**

Please put a √ against all cases where you are CONFIDENT that you would choose the risky treatment.

Please put an X against all cases where you are CONFIDENT that you would REJECT the treatment and accept the certain health state.

Please put a = against the case where you think it would be most difficult to choose between having the risky treatment and not having the treatment

Outcome of treatment:

| Chances of success | | | Chances of failure | | | |
|---|---|---|---|---|---|---|
| 100 | in | 100* | 0 | in | 100* | |
| 99 | in | 100* | 1 | in | 100* | |
| 98 | in | 100 | 2 | in | 100 | |
| 97 | in | 100 | 3 | in | 100 | |
| 96 | in | 100 | 4 | in | 100 | |
| 95 | in | 100 | 5 | in | 100 | |
| 90 | in | 100 | 10 | in | 100 | |
| 85 | in | 100 | 15 | in | 100 | |
| 80 | in | 100 | 20 | in | 100 | |
| 75 | in | 100 | 25 | in | 100 | |
| 70 | in | 100 | 30 | in | 100 | |
| 60 | in | 100 | 40 | in | 100 | |
| 50 | in | 100 | 50 | in | 100 | |
| 40 | in | 100 | 60 | in | 100 | |
| 30 | in | 100 | 70 | in | 100 | |
| 20 | in | 100 | 80 | in | 100 | |
| 10 | in | 100 | 90 | in | 100 | |
| Immediate death preferred | | | | | | |

* You may be willing to accept the treatment but only if it has a chance of failure less than 1 in 100 (i.e. a chance of success which is higher than 99 in 100). If so, at what level of failure would you accept treatment ? .....

**Table 5.1: Health states valued in survey***

| 'Core' (n=3) | 'Rare' (n=33) |
|---|---|
| 111111 | 111112 |
| 124143 | 111323 |
| 623424 | 122424 |
| | 211211 |
| **'Common' (n=22)** | 211212 |
| 111212 | 211222 |
| 111311 | 211223 |
| 111312 | 211442 |
| 211111 | 212222 |
| 222432 | 222222 |
| 224244 | 223423 |
| 311211 | 311212 |
| 311222 | 311422 |
| 313333 | 321412 |
| 323422 | 322313 |
| 322323 | 323333 |
| 422413 | 323433 |
| 422434 | 323435 |
| 423122 | 324434 |
| 521412 | 411412 |
| 523111 | 422334 |
| 525112 | 422533 |
| 525555 | 423423 |
| 624415 | 424425 |
| 624645 | 424444 |
| 625555 | 424524 |
| 625655 | 523421 |
| | 623322 |
| | 623545 |
| | 624422 |
| | 624424 |
| | 624525 |
| | 624534 |

* Where the health states are described by a six digit number indicating the dimension level. Therefore 124143 indicates the following health state: physical functioning at level 1, role limitation at level 2, social functioning at level 4, pain at level 1, mental health at level 4, and vitality at level 3.

**Table 5.2: The distribution of levels of the SF-6D health states included in the valuation survey**

| | Physical | Role | Social | Pain | Mental Health | Vitality |
|---|---|---|---|---|---|---|
| Level | n | n | n | n | n | n |
| 1 | 8 | 19 | 19 | 7 | 17 | 7 |
| 2 | 11 | 39 | 10 | 11 | 22 | 20 |
| 3 | 13 | | 13 | 9 | 10 | 12 |
| 4 | 10 | | 12 | 22 | 6 | 10 |
| 5 | 5 | | 4 | 7 | 3 | 9 |
| 6 | 11 | | | 2 | | |

# Chapter 6

# Results of the Valuation Survey

This chapter reports the results of the survey to value the sample of 58 SF-6D health states. It includes the response rate to the survey, the background characteristics of the respondents, levels of completion of the VAS and SG tasks, descriptive statistics of the valuations (including mean and median by health state, standard deviations, standard errors, and skewness), and tests of the consistency and reliability of the valuations. The changes made in order to prepare the data for the multivariate analysis are also reported. The results are then compared to the MVH valuation surveys that used the same versions of the VAS and SG questionnaires. Finally, the implications for the econometric analyses presented in Chapters 7 and 8 are discussed.

## 6.1 Response rate

One hundred and seventy one people were approached to participate in this study. Fifty five of the respondents were recruited in hospital outpatients clinics and these constitute the patient group. The rest formed the non-patient group and were either working for the NHS or were undergraduate students. Two students refused to participate and four doctors failed to return their questionnaires, resulting in a response rate of 96%. The response of people approached to take part in the survey was better than usually achieved in valuation surveys (MVH, 1994; Jones-Lee et al., 1993), though this was probably due to the use of a convenience sample.

## 6.2 Background characteristics of sample

The age of respondents was between 18-79 years old, with a mean of 40 years (SD=17), and 49% were female (Table 6.1). The majority had non-manual occupations or were students, and a higher proportion had completed their education over the age of 19 ( i.e. 42%) than is found in a random sample of the UK population (OPCS, 1990). The proportion with long-standing illness or disability was nearly 50%. Self-reported health was distributed across all levels of the SF-6D, though as expected the most severe levels were rare.

This sample of respondents was not intended to be representative of the UK general population nor any particular group. Nonetheless, all the key socio-demographic groups used in past valuation studies were represented. The major omission was the absence of people from the manual employed or the unemployed groups.

## 6.3 Completion

There were 1747 VAS completed values out of a potential total of 1760 (i.e. 110 x 12 for non-patients and 55 x 8 for patients), representing a completion rate in excess of 99% (Table 6.2). There was little difference in completion between patients and non-patients (98.9% Vs. 99.4% respectively). These excellent results were achieved without any checking by the supervisors.

Out of a total of 1705 potential SG questions (110 x 12 for non-patients and 55 x 7 for patients)[1] at least one of the three responses asked was completed in 1677 (98.3%) cases. The completion rates, however, varied substantially between the three types of responses requested in the SG question. Respondents indicated the level at which they would accept treatment in 1618 (94.9%) gambles, the level at which they would not accept treatment in 1569 (92.0% gambles), and indicated where it was most difficult to choose (i.e. equal) in only 1149 (67.4%) gambles. There were 73 cases where the respondent indicated he/she would only accept treatment if there was a chance of failure of less than one in 100. The completion rates of patients were significantly lower than in non-patients: 92.2% vs. 95.7%, 87.8% vs. 93.3%, and 41.6% vs. 74.9% respectively by type of SG response.

Respondents were unwilling to identify a point where they would find it most difficult to choose between a chronic state and treatment. However, the numbers who were, indicated the probability of success at which they would not accept the treatment, and the probability at which they would was considerably better. Jones-Lee and his colleagues (1993) found the same problem using an earlier version of the same self-completed SG questionnaire. They recommended using the mid-point between the

[1] Note there are fewer SG than VAS questions for patients. Perfect health (i.e. state 111111) and death are not valued by SG, making two fewer states to value, but patients completed one extra and non-patients two extra gambles in order to compare direct with indirect SG valuations (see below).

probability values where the respondent would accept treatment and point where he/she would not accept the treatment as a proxy for the point of indifference. This extrapolation increased the number of useable SG responses to 1567 and hence to a satisfactory completion rate of 91.9%. This also reduced the disparity in completion rates between patients and non-patient groups (i.e. to 84.2% and 94.2%).

To test the validity of this method of extrapolation, the mid-point was also calculated for those who had indicated a probability value for 'most difficult to choose'. The mid-point value was found to be highly correlated with the actual value (0.98).

### 6.4 Visual analogue scale results

Descriptive statistics are presented for the common health states, that is those with more than 20 observations, on Table 6.3a. The more 'rare' states, those valued with fewer than 20 observations, are presented on Table 6.3b.

The mean VAS ratings for death and health state 111111 were 3.9 and 93.3 respectively. None of the common health states had a mean or median value less than death. In only 16 out of a total of 1747 observations were any of the health states regarded as worse than death. The mean and median health state values were well spread across the VAS scale, and included all deciles. The values are related to the severity of the health, and show a steady fall as severity increases.

The differences in the average values between health states must be interpreted with caution. The standard errors around the means suggest the 95% confidence interval (excluding death and state 111111) to be within the range of $\pm$2.6 to $\pm$7.2 (based on two standard errors either side). Standard errors are related to sample size, and to obtain more precise estimates would require sample sizes of 150 to 400 (depending on the health states) for confidence interval of $\pm$2 points, or sample sizes of 20 to 65 for $\pm$5 points. Furthermore, differences may be owing to the respondent, rather than the content of the health states. (It will be necessary to allow for this respondent effect in the multivariate analysis in Chapter 8.)

The scope for comparing health state values by respondent background characteristics is limited because the number of observations in each sub-group within a health state is usually too small. Only multivariate analysis can disentangle the impact of the SF-6D from the background characteristics of respondents, and this has been undertaken in Chapter 8. However, the numbers of patients and non-patient observations are sufficient for a comparison within health state. There were large differences between patients and non-patients in their average health state values, but there seemed to be no systematic pattern. (i.e. The patient group had values below those of the non-patient group in nine of the 18 health states and the difference was reversed for the remainder (Table 7.3c)).

The mean differed from the median for all states, and this was reflected in a significant level of skewness in distribution. The median was higher for 11 states and lower for 13 (as marked by an asterix on Table 6.3a). It can be seen from Table 6.3a that there was a tendency for the median to exceed the mean value for the milder states (e.g. 111212 and 1113111), and vice versa for the more severe states (e.g. 525555 and 623424). This implies a positive skew in the distribution of the values of the less severe states and a negative skew for the more severe states (Figure 6.1). These distributions result in size of the standard deviations being related to health state severity, with lower values at either extreme, and a peak in the range of the moderate states (e.g. 313333 and 224244 have SDs of 21.56 and 20.73 respectively). This has implications for the multivariate analysis presented in chapter 8.

A closer inspection of the frequency distributions of the health states revealed a multi-modal pattern that indicates a preference for the digits 5 and 10 on the zero to 100 of the scale.

## 6.5 Standard Gamble results

The results from the SG valuations are shown separately for non-patients and patients since they undertook different gambles. Those for non-patients have been divided into the common health states (Table 6.4a) and rare health states (Table 6.4b) as before. The results for gambles with non-fatal treatment outcomes are presented on a third table

(Table 6.4c). The results for the patients are shown on Table 6.4d. There is a further table for both groups showing the distribution of values for those who chose a probability of success above 99 out of 100 (Table 6.4e).

The mean health state values only appear in the upper half of the scale i.e. 50.46 to 98.76 for non-patients and 56.99 to 87.04 for patients. The values from the patient group are not comparable with the non-patient results since the gambles involved different reference states. However, both sets of results show similar patterns, with mean and median health state values declining with increasing severity in the dimensions of the health states.

In just two of the 1638 gambles respondents indicated they would prefer immediate death. At the other end of the scale, on 82 occasions (i.e. 6.6%) individuals were willing to accept the treatment if it had a 99% chance of success or more. On these occasions, the respondents indicated they would have accepted the treatment with probabilities of success between 99.5% to 99.99999%.

There were only three occasions when individuals placed a cross against 100%. This may underestimate the number who did not wish to take a risk, since there were 40 respondents who ticked 100% and crossed 99% without indicating a value between these points at which they would gamble. The procedure recommended by Jones-Lee et al. (1993) assigns a probability value of 99.5% to these individuals, but this may be wrong. They could be indicating that they do not wish to take on risk. Given there were only 40 (i.e. 2.4%) of these cases, assuming values of 100% or excluding them entirely is unlikely significantly to alter the results. Nonetheless, a sensitivity analysis has been performed in the multivariate models presented in Chapter 8 with and without these cases.

A criticism of using SG to value health states has been the view that people are reluctant to take risk at all (Froberg and Kane, 1989b). This would make it very insensitive to differences in the severity of the health states. Froberg and Kane (1989b) cite prospect theory, which suggests people weight losses (i.e. risk of death) more than gains (i.e.

return to perfect health). The results from this survey suggest people are prepared to take risks, and sometimes large risks of death to cure health problems. This result is particularly striking given the fact that many of the states defined by the SF-6D were comparatively mild. People may be risk averse, but this does not prevent them from being willing to take risk. Furthermore, the amount of risk people are willing to take is related to the severity of health problems.

The risk respondents were prepared to take were often quite low, and typically between 95% to 100%. However, this may also indicate that the probability categories offered to respondents were insensitive at this end of the scale. This was the reason for allowing respondents to opt for choosing their own probability value and 5% did so. Of course, there may have been some inertia to 'opt out' of the scale, and so some respondents may have overstated the degree of risk they were willing to take. There could be a case for increasing a further the number of categories at the end of the scale and/or making it easier to 'opt out'.

The frequency distributions of the values by health state were positively skewed, and the median exceeded the mean value for all except two of the health states for non-patients (Table 6.4a). This skewness reflected the fact that all mean and median SG values are in the top half of the scale. There is also a positive association between standard deviation and the severity of the health state in the non-patient group. (i.e. the highest SDs are for health states 52555 and 625555, and these have the lowest mean value). For patients, this inverse association between standard deviation and the mean was much weaker. The overall frequency distribution of responses were multi-modal, as for VAS, but this was a result of the response categories in the SG question rather than digit preference.

There was considerable dispersion in health state values. Non-patient SG mean values had standard errors within the range 0.48-12.78, and they increased with health state severity. The milder states had more precise estimates and hence correspondingly require smaller sample sizes. To obtain a 95% CI of +2 points would require between

100-500 observations. The more severe states would require up to 900 observations to achieve the same level of precision.

### 6.6 Reliability

A reliable technique for eliciting preferences should reproduce the same valuation of a health state at different points in time. This is usually tested by administering the question to the same individual at two points in time, typically at an interval of a week or more in order to avoid the risk of the respondent remembering his/her previous answer (Streiner and Norman, 1989). In this survey, one health state (124143) has been administered twice within one sitting, allowing a form of re-test reliability known as an internal or split test reliability (Torrance, 1976). There is a risk of respondents remembering, but the health states defined by the SF-6D are complex and respondents are being asked to value a large number of them. Furthermore, the health states appeared in different envelopes for the VAS exercise and were four gambles apart in the SG task. Split test reliability was assessed in terms of a Spearman rank correlation and the mean differences between tests.

The test was undertaken by 33 non-patients on the VAS and 19 on SG. The coefficients of reliability, as measured by the Spearman rank correlation coefficient, were 0.46 and 0.64 for the VAS and SG respectively. The mean difference between VAS values at test and re-test of 0.61 (SD=18.47) was not statistically significant. The mean valuation for SG was significantly higher at the second administration by 8.25 (SD=16.60, $\alpha$ = 0.034).

This limited evidence suggests a significant degree of instability in respondents' valuations by either method. The instability of the VAS valuations seems to be unbiased, but the evidence of bias in the SG valuation is potentially more of a problem. It could be explained by the relative position of the two assessments in the order of the SG questions. Health state 143123 was the third and sixth of the SG questions. The first two questions concerned severe health states (e.g. 52555), whereas questions four and five were comparatively mild states (e.g. 211111), and this may have influenced the respondent's answer. After the two questions with mild states, the respondent may have

become more accustomed to using the upper part of the scale and hence chose a higher value. The evidence is, however, too limited to draw any firm conclusions.

### 6.7 Consistency with SF-6D

An important check on the extent to which respondents were able to understand the valuation tasks is to examine the consistency of their responses with the health state classification (e.g.MVH, 1994). For many pairs of health states defined by the SF-6D, one state can be regarded as dominant over the other if it has a less severe health problem on one or more dimensions but not worse health problem on the remaining dimensions. In this case, the dominant health state should be either logically preferred or regarded as equal to the other state. Respondents who understand the valuation task and the SF-6D description, for example, should <u>not</u> prefer health state 311212 to 311211. The degree of logical consistency is described in Tables 6.5a and 6.5b using the classification of strict consistency with the predetermined rank (i.e. >), strict inconsistency in the case of reversals (i.e. <), and equality (i.e.=). Such prior judgements are not possible for all paired comparisons, since many involve trade-offs between dimensions (e.g. 124143 versus 523112).

In paired comparisons of the mean and median values of the common health states there were no logical inconsistencies in VAS data (Table 6.3a), and only two in the SG data (Table 6.4a and 6.4d). The SG inconsistencies were between 523111 and 523112 (i.e. 90.2 vs. 90.8) and between 311222 and 313333 (i.e. 73.0 and 74.6). These may have arisen from the fact that these states were valued by different respondents. A better test of consistency is undertaken at the individual level.

The level of consistency has been examined by respondents for 40 pairs of states for the VAS data and for 23 pairs for SG in non-patients, and 28 pairs of states for VAS and 14 for SG in patients (Tables 6.5a and 6.5b). There were 94.6% of pairs of health states where the VAS valuations were strictly consistent with the logical ordering (i.e. 1537 out of 1626), 1% were equal, and 4.3% were strictly inconsistent. The level of consistency was significantly lower amongst patients where the proportions were 88.2%, 0.5% and 11.3% respectively against 97.1%, 1.4% and 1.6% for non-patient

respondents. The proportion of strictly consistent pairs of SG valuations was less at 83.9%, with 7.9% being equal, and 8.2% inconsistent. The level of consistency was lower again for patients (i.e. 65.5%, 16.4% and 18.1% by category).

It is important to establish whether this inconsistency is random. This is done by examining the relationship between the level of inconsistency and the difference between states as measured by a distance score. This score is calculated as the sum of the differences between each level of the two states (i.e. the distance scores between states 111312 and 211111 is 1 + 0 + 0 + 2 + 0 + 1 = 4). Consistency was found to be positively correlated with the difference between the state as measured by the distance score for values (Table 6.7). Most VAS inconsistencies occurred between health states with distance scores of three or less. For SG, inconsistencies occurred for distance scores extended up to 10. The proportion of strictly consistent responses exceeded the inconsistent responses for VAS and SG values for all pairs.

An important aspect of data quality is whether respondents appear to understand the task. This has been indicated by the consistency of respondents' health state valuations with the logic of the SF-6D. The levels of inconsistency reported in the MVH survey were 2.5% for VAS and 6.2% for TTO (MVH, 1994) and compare favourably with the 4.3% and 8.2% respectively found in this study. Furthermore, these inconsistencies were not random, but related to the size of the difference between the health states. This supports the view that most respondents were able to understand the descriptions of the SF-6D and the valuation tasks.

## 6.8 Consistency between direct and indirect SG valuations

For reasons described in the last chapter, most of the gambles undertaken by patients, and of the 12 undertaken by non-patients, have a non-fatal outcome for the worst reference state. The original aim was to pool the results of these non-fatal outcome gambles with the conventional gambles involving perfect health and death as reference outcomes. According to the axioms of expected utility theory, non-fatal gambles can be transformed onto the full health-death scale when the non-fatal reference state has been

valued in a full health and death gamble. In the SG question, the respondent is asked to indicate the probability of success P, such that:

$$U_x = P\,(U_y) + (I\text{-}P)\,(U_z) \qquad (1)$$

where the best outcome is $U_y$ and the worst outcome is $U_z$. When $U_y$ is full health and $U_z$ is death, these are arbitrarily set to one and zero, and hence $U_x = P$. Changing the worst outcome to a non-fatal state requires the value of $U_z$ to be obtained from another gamble with outcomes of perfect health and death, in order to derive a value for $U_x$ on the death to full health scale. Supposing $U_z = P_z$, then $U_x$ becomes:

$$U_x = P\,(1) + (1 - P)\,(P_z) \qquad (2)$$

The values for $U_x$ obtained in this chain of two gambles should be the same as those obtained directly in a single gamble involving the reference states full health and death.

It is important to demonstrate this prediction before pooling the SG data sets. In this survey, patients valued five health states in gambles with a non-fatal outcome as the worst reference state (i.e. $U_z$). They had previously valued the reference health state in a conventional gamble (i.e. with full health and death as the reference states), and therefore an indirect valuation was obtained for the five states on the full health (i.e. 1) to death (i.e. 0) scale. Non-patients also undertook two non-fatal gambles involving a worse reference state valued earlier in a conventional gamble. There are nine indirect values of health states (i.e. obtained from chains of two gambles) for which there were also direct valuations. The indirect values were found to exceed the direct estimates in all comparisons at the 1% level (Table 6.8).

The results from comparing direct with indirect values indicate a substantial departure form the predictions of EUT. Similar results have been obtained in studies by Llewellyn-Thomas et al., 1982 and Jones-Lee et al., 1993; and Rutten-von Molken et al., 1994). To be consistent, respondents should have been prepared to take substantially greater risk in gambles involving a non-fatal treatment failure. The implication is that

health states obtained from gambles using different reference states should not be pooled. This implies that most of the SG values obtained from patients should be excluded. To avoid such a loss of data in the modelling, it is worth at least considering the possibility of a systematic explanation for these differences which could be used to model a relationship between these estimates.

Llewellyn-Thomas et al. (1982) examined the consequences of substituting the reference states (i.e. full health as well as death). The change in the best outcome produced similar mean utilities, and statistically significant differences in only three out of eight comparisons. However, substituting for death had a much larger impact, and was significant for all eight of the comparisons. The strong influence of the failure outcome led the authors to suggest there was a 'framing effect', with respondents regarding the potential outcomes of death as a loss but outcomes without death as gains. According to prospect theory individuals would be risk-seeking over potential losses and risk-averse at the prospect of gains (Kahneman and Tversky, 1979). The probabilities of success from gambles involving death would be lowered by risk-seeking and conversely raised for non-fatal gambles by risk aversion. This would seem an implausible hypothesis, since in an SG question the respondent is being asked to imagine he/she is in a health state which is better than the worst reference state (i.e. $U_x > U_z$). There is a possibility of very ill respondents regarding treatment failure as better than their actual state, and this may have been true for the patients in the study by Llewellyn-Thomas and colleagues who were receiving radiotherapy, but this is unlikely to be the case for the respondents to the Sheffield valuation survey. In eight of the nine comparisons presented on Table 10, the valuations were provided by non-patients and the majority were in good or excellent health.

A more likely explanation has been proposed by Morrison (1994) based on a general aversion to gambling, known as the 'gambling effect' (Gafni, 1994; Bombadier et al., 1982). In the context of the SG question used in this study, there could also be a general aversion to having a risky treatment such as surgery. Morrison proposes a disaggregation of the probability of success from each gamble into a generic component for the gambling effect and a specific component reflecting a person's strength of

preference for the health state. The generic component could be constant between gambles. The consequence of this model is to replace $U_i$ (i relates to health states x, y or z) by $\lambda^i_0 + \lambda^i$, where $\lambda_0$ is the generic component and $\lambda_1$ the specific component. The consequences for equations (1) and (2) are as follows:

$$\text{Direct } U_x = \lambda^x_0 + \lambda^x_1 \qquad (1a)$$

$$\text{Indirect } U_x = \lambda^y_0 + \lambda^y_1 + (1 - \lambda^x_0 - \lambda^x_1) * (\lambda^z_0 + \lambda^z_1) \qquad (2a)$$

Morrison (1994) has developed a method of correcting for this bias by estimating the generic component $\lambda^i_0$. In her work, this component was estimated for each SG comparison by regressing the probability of success of one gamble onto another with the same or a similar outcome failure. The generic component is then removed from the SG data and the direct and indirect mean health state utilities re-estimated.

Undertaking such an adjustment procedure may reduce some of the discrepancies in Table 6.8, but it cannot eliminate all of them. The original values of the non-fatal gambles are already higher than the direct value in five of the nine comparisons. This degree of inconsistency suggests many, if not most of the respondents, were unable to distinguish between gambles with different treatment failures. It is the prospect of a failure in the treatment which seems to be the focus of attention rather than its consequences. As observed by Jones-Lee et al. (1993) in their studies of injury state valuation *"..Varying the severity of the consequence of failure in the risky treatment requires a conceptually more difficult adjustment than keeping the risky treatment the same and varying the severity of the injury description"* (P62). Given this level of confusion, it was decided to exclude all gambles involving non-fatal treatment failures, and hence all patient valuations.

The failure of indirect valuation methods to yield results comparable with direct methods adds to the large and growing body of evidence of EUT violations (Shoemaker, 1982, Appleby and Stammer, 1987). It may have been due to respondent confusion, rather than alternative behavioural explanations to EUT, such as suggested by prospect

theory or gambling aversion. The general implication is that the outcome of treatment failure should not be varied between gambles intended to generate a measure reflecting preferences on the same cardinal scale.

## 6.9 Data Preparation

The data has been prepared for the multivariate analysis in two ways. The first has been to transform the VAS ratings onto a full health to death scale using equation 3.1. This permits the comparison of VAS scores between respondents. The second has been to remove respondents who had major inconsistencies in their answers.

Previous studies have 'cleaned' health state valuation data by eliminating respondents who were thought to have been confused by the valuation task (Kaplan et al., 1979; Torrance, 1982; Dolan et al., 1994), where confusion has been defined as major inconsistencies in the values obtained from respondents. Excluding such cases has the advantage of improving the precision of the estimates since it tends to reduce the level of variance in the data set. The disadvantages are that it reduces the size and the representativeness of the data set. Inconsistencies have been found to be higher amongst patients in this study. Furthermore, care must be taken when interpreting inconsistencies as evidence of confusion. Inconsistencies may represent simple one-off mistakes and some may reflect departures from the axioms of EUT or consumer choice theory more generally (Loomes, 1993). Nonetheless, it is usual to exclude the most extreme cases of inconsistency with a health scale, and this has been done in preparation for the multivariate analysis.

Five respondents (all patients) were excluded from the VAS data set because they had valued the majority of the health states more highly than the best health state (i.e. state 111111), indicating they did not understand the task (this is similar to the criterion used by Kaplan et al., 1979; Torrance, 1982, and MVH, 1994). All other inconsistencies have been retained. More observations were excluded from the VAS data set as a result of the application of equation 3.1 ( i.e. setting perfect health to one and death to zero). It resulted in the exclusion of a further four respondents who did not value the best health

state (111111) and three who failed to value death, since without these it is not possible to calculate an adjusted score.

For SG, one respondent has been excluded for placing crosses above the ticks. No respondents were excluded for refusing to take any risk. The main cause of exclusion has been gambles involving reference states of perfect health with a non-fatal outcome since the results presented do not support the prediction of EUT that indirect estimates are equivalent to those obtained directly. This has resulted in all patients being from the multivariate analysis, since it leaves only two observations per patient, and this was not regarded as sufficient to adjust for the respondent effect (see Chapter 8).

The consequences of these changes have been as follows. There were 165 respondents who provided 1582 VAS ratings (excluding state 111111 and death) and 1567 SG values for 58 health states (Table 6.2). After adjusting the VAS ratings and excluding respondents for major inconsistencies, there were a total of 1357 VAS observations by 155 respondents. This represents a useable rate amongst the responses of 89.8%[2]. The adjusted VAS ratings are presented by health state on Table 6.9. The exclusion of all patients and one non-patient for gross inconsistencies left 1037 SG observations from 106 respondents. This amounts to an overall useable rate of 60.8% amongst all responses, 78.6% in the non-patient group amongst all their responses, and 94% in the non-patient group for responses to the gambles with the fatal reference. These are the data sets used in the econometric analyses in Chapters 7 and 8.

## 6.10 Implications of distributions in health state values for multivariate analysis

The distributions of VAS values were found to change from positively to negatively skewed with increasing health state severity. This result has also been found in the MVH survey using the same instrument to value the EQ-5D. This may have been an artefact of the scale caused by the end points acting as 'ceilings' and 'floors' to the

[2] The useable rate is calculated as the potential number of observations from the returned questionnaires divided by the number after exclusions for inconsistency and adjustment.

distribution. This explanation would be consistent with the tendency for larger SD values to be associated with the mid-range health states.

The distributions of the SG values within health states were found to be positively skewed for all health states. This skewness may also have been an artefact of the scale, a view confirmed by the positive association between SD and health state severity. For SG a more likely explanation is that the concentration of observations around 95% to 100% may reflect peoples preferences.

The skewness of the distributions raises problems with the mean as a measure of central tendency. It results in mean values being heavily influenced by a few observations in the tail of the distribution. For example, three mild health states with median SG values of 99.0 or above have mean values of 96.2, 96.6 and 98.8 (Table 6.4a). The mean values imply people are prepared to take a risk of death of up to one in 26 to cure these mild conditions, yet these values do not reflect the values of the majority of respondents. The extremist therefore has more leverage to influence the value of these mild states than respondents at the mode.

One approach to dealing with the influence of a few outliers would be to trim the tails of the distributions. The grounds for doing this are either that people do not correctly understand the task, or that since they hold such extreme views they should be excluded. The former is a common argument, but a better approach is to remove only observations shown to be grossly inconsistent. It would also seem to be unfair to exclude someone simply because their views did not correspond to those of the majority. Another solution, often favoured by statisticians, would be to use the median as a measure of central tendency.

The choice of measure of central tendency is not only a statistical issue. The mean is conventionally used for determining efficiency since this best reflects the strength of people's preferences. In a market, a person's willingness to pay is not merely a vote in favour of one commodity over another but an expression of the intensity with which a person holds such a view. This extends into cost-benefit analysis, where according to

the compensation principle the gainers can be allowed to compensate the losers (Friedman, 1984). The mean is therefore preferred for assessing economic efficiency (Drummond et al., 1987).

The use of the median can be justified in terms of the median voter argument. In a democracy, extremists are not given more weight than moderates because of the intensity of their views. The purpose of economic evaluation is to assist in the making of public decisions, and on these grounds it could be argued that the median more closely reflects the political system. This debate has not been resolved in the literature and therefore both the mean and median average will be used in the multivariate analysis.

The existence of multiple modes in the VAS distributions can be explained by digit preference. This phenomena has been observed using the same VAS instrument with the EQ-6D (Parkin, 1991). Parkin (1991) argued the VAS thermometer was too finely graded and that the values should be rounded to the nearest quintal or even decimal scores. He rounded a set of EQ data to the nearest quintal, but found there was little effect on the means, and though the distribution became smoother, some multiple modes remained. Parkin (1991) has also argued that the mean was too exact given such apparent discreteness in the data, and argues for the use of medians or ranges of quintal or decimal scores.

The SG values were also found to be discontinuous owing to the discrete scale of responses (at least up to 95 out of 100), and hence a similar argument might apply. However, the intervals between the discrete values of the SG and the preferred VAS digit also express an intensity of preference and as already argued, should not be ignored in the assessment of benefit for economic evaluation.

The existence of heterogeneity in the variance of health state values also has important implications for the statistical analysis. Standard Ordinary Least Squares regression would not be appropriate without some adjustment to the distribution of the data.

## 6.11 Comparison with the MVH Survey

The two surveys have used the same versions of the preference elicitation techniques and hence there is an opportunity to compare the quality of the data in terms of completeness and consistency. This should provide some insight into the question of whether the larger size of the SF-6D compared to the EQ-5D has been at the expense of respondent acceptability and comprehension. It may also indicate the extent to which the classifications overlap in terms of health state values.

The high levels of response and completion achieved in the Sheffield valuation survey compared favourably with the results of the MVH surveys. The levels of inconsistency were also comparable with the results of the MVH survey. It might be concluded, therefore, that the extra size and complexity of the SF-6D was not at the expense of the respondent comprehension. However, there are important differences between the surveys. Respondents in the MVH survey were seen individually by interviewers and this would have been an advantage over the self-completion in groups. On the other hand, the Sheffield sample was not representative of the general population and had a lower proportion in manual occupations and fewer people who left school at 16 or earlier. These characteristics are likely to have improved the consistency levels of the Sheffield survey. Finally, it is doubtful whether the consistency results can be compared since the results are influenced by the selection of states since those pairs of states that are further apart generate fewer inconsistencies. Nonetheless, for the samples of states used in the surveys the findings of this survey are promising for the SF-6D.

Evidence of the descriptive validity of the SF-6D and EQ-5D (Chapters 3 and 4) suggests that SF-6D health states will be more concentrated at the milder end of the health spectrum than those of EQ-5D. The VAS results confirm this by showing a larger proportion of SF-6D health states at the upper end of the distribution than EQ-5D states in the MVH survey (Table 6.10). The difference was less than might be expected: 35% of SF-6D states had mean values in excess of 60 against 29% of EQ-5D states, and 34% of SF-6D states have values below 40 compared to 51% of EQ states. This may reflect the doubts about the validity of the VAS scale in making interpersonal comparisons (Nord, 1993; Richardson, 1994). There is a tendency, for example, for respondents to

use the entire length of the scale, regardless of what is being valued (Stevens and Valenter, 1957). This 'spreading effect' is likely to increase the value of EQ-5D health states since they will be pushed up the scale by the nastier states. Conversely, SF-6D states will be pushed down the scale by the nicer states. This would dampen the real differences existing between EQ-5D and SF-6D.

The SG valuations of the SF-6D and the published TTO valuations of the EQ-5D cannot be formally compared. However, it is interesting to note that there were no mean SG values for SF-6D health states below 50, yet over half of the mean TTO values were below this point. A more telling result was that 38% of TTO values were negative values, indicating they were worse than death. None of the SF-6D states were regarded as worse than death. Furthermore, 70% of SF-6D health states had mean values above 80 compared to 12% for EQ-5D. This evidence provides further support for sensitivity of SF-6D across milder health problems, but suggests it would not be suitable for patients with very severe health problems

## 6.12 Conclusion

The valuation survey has been successful in achieving its primary objective of generating VAS and SG data sets for estimating preference weights for the SF-6D using statistical methods.

There was a good response amongst those approached to participate in the survey. The respondents were not a representative sample of any one group, but nonetheless reflected a range of backgrounds and illness experiences. The quality of the VAS and SG data in terms of the levels of completion and consistency compared favourably with other surveys. There was evidence of instability in the valuations from the split test. This instability and the confounding effect of the respondent on the health state valuations made formal comparisons of average health state values inappropriate.

The data sets were prepared for the multivariate analysis by removing major inconsistencies from each and transforming the VAS data. The most important exclusion was the patients' SG valuations owing to evidence of inconsistency between

direct health state valuations and those obtained indirectly via chains of gambles. Otherwise the amount of data lost was minimal (i.e. 10.1% for VAS and 6% for SG).

The distributions of the VAS and SG valuations raised questions about the appropriate measure of central tendency and the methods of multivariate analysis. The mean is usually the preferred measure of central tendency for economic evaluation but there are strong arguments for the median and therefore both are modelled in the multivariate analysis. The relationship between variance and the severity of health state suggests ordinary least squares will be an inappropriate technique for estimating the relationship between SF-36D and the valuations.

**Table 6.1: Characteristics of respondents in the valuation survey (n=165)**

**1) Age and Sex**

| Age Groups (years) | Males n | % | Females n | % |
|---|---|---|---|---|
| 16-24 | 14 | 9 | 11 | 7 |
| 25-44 | 30 | 18 | 46 | 28 |
| 45-64 | 24 | 15 | 17 | 10 |
| 65-79 | 14 | 9 | 7 | 4 |

**2) Age finishing education**

| | n |
|---|---|
| Under 17 | 57 |
| 17-18 | 23 |
| Over 19 | 57 |
| Missing | 28 |

**3) Occupation**

| | n |
|---|---|
| Professional | 32 |
| Managerial | 51 |
| Other non-manual | 12 |
| Skilled manual | 2 |
| Semi-skilled manual | 2 |
| Retired or student | 59 |
| Missing | 7 |

**4) General Health**

| | n | % |
|---|---|---|
| Excellent | 36 | 22 |
| Very good | 61 | 37 |
| Good | 29 | 18 |
| Fair | 28 | 17 |
| Poor | 9 | 5 |
| Missing | 2 | 1 |

**5) Long-standing illness, disability or infirmity**

| | n | % |
|---|---|---|
| Yes | 80 | 49 |
| No | 83 | 50 |
| Missing | 2 | 1 |

**6) Distribution of self-reported health across the SF-6D**

| Physical Functioning | | Role Limitation | | Social Functioning | |
|---|---|---|---|---|---|
| Level | n | Level | n | Level | n |
| 1 | 90 | 1 | 121 | 1 | 97 |
| 2 | 42 | 2 | 40 | 2 | 32 |
| 3 | 11 | | | 3 | 15 |
| 4 | 7 | | | 4 | 13 |
| 5 | 13 | | | 5 | 5 |
| 6 | 2 | | | | |

| Bodily Pain | | Mental Health | | Vitality | |
|---|---|---|---|---|---|
| Level | n | Level | n | Level | n |
| 1 | 89 | 1 | 84 | 1 | 45 |
| 2 | 27 | 2 | 69 | 2 | 88 |
| 3 | 19 | 3 | 6 | 3 | 24 |
| 4 | 18 | 4 | 2 | 4 | 4 |
| 5 | 8 | 5 | 1 | 5 | 1 |
| 6 | 2 | | | | |

**Table 6.2: Completion of valuation task**

| | Patient | | Non-patient | | Overall | |
|---|---|---|---|---|---|---|
| | n | % | n | % | n | % |
| VAS Rating | | | | | | |
| Total number | 440 | | 1320 | | 1760 | |
| Completed | | | | | | |
| - Raw data | 435 | 98.9 | 1312 | 99.4 | 1747 | 99.3 |
| - Adjusted data | 336 | 76.4 | 1176 | 89.1 | 1512 | 85.9 |
| Standard Gamble | | | | | | |
| Total number | 385 | | 1320 | | 1705 | |
| Completed | | | | | | |
| Any item | 377 | 97.9 | 1300 | 98.5 | 1677 | 98.3 |
| 'Upper' | 355 | 92.2 | 1263 | 95.7 | 1618 | 94.9 |
| 'Lower' | 338 | 87.8 | 1231 | 93.3 | 1569 | 92.0 |
| 'Best' | 160 | 41.6 | 989 | 74.9 | 1149 | 67.4 |
| Mid-point or 'best' | 324 | 84.2 | 1243 | 94.2 | 1567 | 91.9 |

**Table 6.3a: Unadjusted VAS ratings - 'core' and 'common' health states**

| Health State | Mean | SD | SE | Median | I-Q range | n |
|---|---|---|---|---|---|---|
| 111111 | 95.35 | 13.80 | 1.08 | 98.00 | 94-100 | 164 |
| 111212 | 85.89 | 11.50 | 1.94 | 90.0 | 85-90 | 35 |
| 111311 | 84.45 | 12.88 | 1.69 | 90.0 | 80-94 | 58 |
| 111312 | 84.53 | 13.46 | 2.31 | 87.0 | 80-91 | 34 |
| 124143 | 53.96 | 13.96 | 1.34 | 55.0 | 40-68 | 199 |
| 211111 | 83.56 | 12.69 | 2.12 | 85.0 | 76-94 | 36 |
| 222432 | 46.47* | 17.47 | 3.09 | 45.0 | 30-59 | 32 |
| 224244 | 43.79* | 20.73 | 2.55 | 43.0 | 30-60 | 66 |
| 311211 | 68.66 | 14.68 | 2.64 | 70.0 | 60-80 | 31 |
| 311222 | 64.27 | 15.27 | 2.51 | 67.0 | 50-75 | 37 |
| 313333 | 59.74 | 21.56 | 3.65 | 65.0 | 50-75 | 35 |
| 322323 | 61.56 | 17.20 | 2.87 | 65.0 | 51-75 | 36 |
| 323422 | 45.37* | 18.91 | 3.45 | 45.0 | 30-58 | 30 |
| 422413 | 54.89 | 19.69 | 3.33 | 60.0 | 40-70 | 35 |
| 422434 | 26.73* | 11.35 | 2.42 | 25.0 | 19-35 | 22 |
| 423122 | 54.97* | 15.80 | 2.67 | 50.0 | 45-70 | 35 |
| 521412 | 50.23 | 16.78 | 3.58 | 53.0 | 40-61 | 22 |
| 523111 | 63.95* | 15.20 | 3.26 | 62.0 | 50-80 | 22 |
| 525112 | 43.66* | 16.06 | 2.11 | 42.0 | 30-55 | 58 |
| 525555 | 16.82* | 13.01 | 2.77 | 15.0 | 5-26 | 22 |
| 623424 | 34.81* | 15.49 | 1.30 | 32.0 | 22-45 | 143 |
| 624415 | 23.60 | 13.16 | 2.40 | 25.0 | 10-35 | 30 |
| 624645 | 16.32* | 10.55 | 2.25 | 15.0 | 9-21 | 22 |
| 625555 | 8.52 | 6.14 | 2.21 | 15.0 | 5-30 | 31 |
| 625655 | 20.62* | 17.52 | 2.21 | 15.0 | 5-30 | 63 |
| Death | 3.94* | 10.34 | 0.82 | 0.0 | 0-5 | 161 |

Note: * is where the mean value exceeds the median

**Table 6.3b: Unadjusted VAS values continued - 'rare' health states**

| Health State | Mean | SD | n |
|---|---|---|---|
| 111122 | 82.13 | 18.06 | 8 |
| 111323 | 81.11 | 10.24 | 9 |
| 122424 | 54.11 | 11.97 | 9 |
| 211211 | 86.78 | 7.03 | 9 |
| 211212 | 80.50 | 11.70 | 8 |
| 211222 | 74.89 | 13.79 | 9 |
| 211223 | 63.44 | 21.24 | 9 |
| 211442 | 56.38 | 18.23 | 8 |
| 212222 | 72.38 | 10.31 | 8 |
| 222222 | 65.67 | 17.90 | 9 |
| 223423 | 44.25 | 25.89 | 8 |
| 311212 | 66.63 | 16.07 | 8 |
| 311422 | 60.67 | 14.59 | 9 |
| 321412 | 62.13 | 11.93 | 8 |
| 322313 | 55.22 | 16.02 | 9 |
| 323333 | 53.60 | 15.92 | 10 |
| 323433 | 37.30 | 19.87 | 10 |
| 323435 | 36.25 | 8.07 | 8 |
| 324434 | 44.88 | 15.27 | 8 |
| 411412 | 70.88 | 16.69 | 8 |
| 422334 | 43.44 | 19.89 | 9 |
| 422533 | 32.22 | 12.02 | 9 |
| 423423 | 38.56 | 18.57 | 9 |
| 424425 | 35.50 | 12.47 | 8 |
| 424444 | 44.89 | 16.56 | 9 |
| 424524 | 34.88 | 21.71 | 8 |
| 523421 | 46.00 | 13.78 | 10 |
| 623322 | 35.75 | 24.26 | 8 |
| 623545 | 14.33 | 6.63 | 9 |
| 624422 | 30.00 | 14.02 | 9 |
| 624424 | 20.75 | 8.81 | 8 |
| 624525 | 24.44 | 10.98 | 9 |
| 624534 | 17.22 | 12.77 | 9 |

**Table 6.3c: Unadjusted VAS ratings for patients and non-patients - 'common' health states**

| Health State | Patient (n=55) | | | | Non-patient (n=110) | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Median | n | Mean | SD | Median | n |
| 111111 | 86.83 | 21.71 | 98 | 55 | 96.64 | 4.37 | 98 | 109 |
| 111212 | 79.62 | 14.63 | 90 | 13 | 89.59 | 7.27 | 90 | 22 |
| 111311 | 81.14 | 11.05 | 90 | 14 | 85.50 | 13.36 | 90 | 44 |
| 111312 | 77.25 | 19.23 | 89 | 12 | 88.50 | 6.63 | 89 | 22 |
| 124143 | 57.90 | 19.60 | 52 | 68 | 51.91 | 18.37 | 52 | 131 |
| 211111 | 78.21 | 16.24 | 88 | 14 | 86.96 | 8.60 | 88 | 22 |
| 224244 | 47.31 | 12.73 | 43 | 13 | 42.92 | 22.27 | 43 | 53 |
| 311222 | 60.47 | 19.26 | 72 | 15 | 66.86 | 11.61 | 72 | 22 |
| 313333 | 64.92 | 22.71 | 63 | 13 | 56.68 | 20.77 | 63 | 22 |
| 322323 | 56.64 | 20.90 | 65 | 14 | 64.68 | 14.01 | 65 | 22 |
| 422413 | 45.85 | 21.73 | 60 | 13 | 60.22 | 16.67 | 60 | 22 |
| 423122 | 59.23 | 14.64 | 50 | 13 | 52.45 | 16.23 | 50 | 22 |
| 521412 | 77.25 | 19.23 | 53 | 12 | 50.23 | 16.78 | 53 | 22 |
| 523111 | 47.31 | 12.73 | 63 | 13 | 63.96 | 15.30 | 63 | 22 |
| 525112 | 49.79 | 19.65 | 40 | 14 | 41.70 | 14.46 | 40 | 44 |
| 525555 | 15.16 | 12.80 | 14 | 18 | 16.82 | 13.00 | 14 | 22 |
| 623424 | 35.96 | 17.47 | 15 | 55 | 34.09 | 14.18 | 32 | 88 |
| Death | 2.25 | 4.79 | 0 | 53 | 4.77 | 12.11 | 0 | 108 |

**Table 6.4a: SG results for non-patients - common health states**

| Health State | 'Best'[1]. | | | | | | Min | Max |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | SE | Median | IQ | n | Mean | Mean |
| 111212 | 96.22 | 6.14 | 1.37 | 99.25 | 96.0-99.5 | 20 | 94.50 | 97.74 |
| 111311 | 98.76 | 1.62 | 0.24 | 99.00 | 98.5-99.5 | 44 | 97.44 | 99.59 |
| 111312 | 95.69 | 5.54 | 1.31 | 97.50 | 95.0-99.5 | 18 | 92.94 | 97.81 |
| 124143 | 88.13 | 12.82 | 1.14 | 95.00 | 85-97 | 127 | 83.12 | 91.77 |
| 211111 | 96.61 | 7.78 | 1.70 | 99.00 | 98.0-99.5 | 21 | 95.32 | 97.80 |
| 222432 | 91.98 | 9.42 | 1.69 | 95.50 | 90.0-98.0 | 31 | 87.65 | 94.97 |
| 224244 | 88.14 | 12.99 | 1.84 | 92.75 | 85.0-97.0 | 50 | 83.30 | 91.94 |
| 311211 | 97.14 | 2.33 | 0.43 | 98.00 | 96.0-99.0 | 29 | 95.17 | 98.48 |
| 311222 | 92.18 | 8.90 | 1.90 | 96.00 | 85.00-99.0 | 22 | 88.77 | 94.20 |
| 313333 | 81.46 | 18.08 | 4.26 | 85.00 | 70.0-98.0 | 18 | 75.67 | 85.24 |
| 322323 | 90.73 | 8.62 | 1.84 | 90.00 | 85.0-98.0 | 22 | 85.22 | 95.32 |
| 323422 | 90.98 | 9.61 | 1.76 | 95.50 | 85.0-97.0 | 30 | 86.40 | 94.37 |
| 422413 | 83.05 | 18.20 | 3.8 | 86.25 | 75.0-96.0 | 22 | 79.90 | 88.95 |
| 422434 | 88.68 | 12.66 | 2.70 | 94.00 | 85.0-95.5 | 22 | 83.05 | 92.73 |
| 423122 | 79.50 | 19.92 | 4.70 | 85.00 | 65.0-97.0 | 18 | 73.06 | 84.29 |
| 521412 | 82.61 | 20.47 | 4.82 | 90.00 | 65.0-97.0 | 18 | 77.06 | 85.30 |
| 523111 | 88.81 | 11.94 | 2.55 | 93.75 | 85.0-97.5 | 22 | 84.14 | 92.57 |
| 525112 | 90.36 | 12.51 | 1.98 | 95.75 | 90.0-97.0 | 40 | 85.98 | 93.65 |
| 525555 | 50.46 | 26.06 | 5.56 | 50.00 | 35.0-75.0 | 22 | 45.25 | 60.23 |
| 623424 | 77.58 | 19.08 | 2.11 | 84.00 | 65.0-75.0 | 82 | 70.07 | 81.58 |
| 624415 | 82.97 | 16.07 | 2.93 | 85.00 | 80.0-95.0 | 30 | 76.73 | 87.50 |
| 624645 | 62.17 | 28.56 | 6.39 | 70.00 | 30-80.0 | 20 | 53.20 | 68.43 |
| 625555 | 54.34 | 31.31 | 5.92 | 60.00 | 20-80.0 | 28 | 49.61 | 61.86 |

Note: [1]. 'Best' is the point where it is most difficult to choose or the mid-point between the minimum and maximum (i.e. the point of indifference).

**Table 6.4b: SG results for non-patients continued - 'rare' states**

| Health State | Mean | SD | n |
|---|---|---|---|
| 111122 | 96.81 | 4.28 | 8 |
| 111323 | 98.42 | 1.57 | 9 |
| 122424 | 88.45 | 11.80 | 10 |
| 211211 | 96.50 | 3.86 | 9 |
| 211212 | 95.21 | 4.95 | 7 |
| 211222 | 87.33 | 14.98 | 7 |
| 211223 | 91.95 | 13.43 | 8 |
| 211442 | 90.06 | 11.88 | 8 |
| 212222 | 95.07 | 5.14 | 8 |
| 222222 | 88.45 | 15.34 | 10 |
| 223423 | 84.50 | 15.55 | 8 |
| 311212 | 90.29 | 17.80 | 7 |
| 311422 | 86.07 | 11.54 | 7 |
| 321412 | 90.56 | 7.64 | 8 |
| 322313 | 89.72 | 15.54 | 9 |
| 323333 | 86.31 | 14.54 | 8 |
| 323433 | 85.38 | 14.95 | 8 |
| 323435 | 78.83 | 23.09 | 9 |
| 324434 | 76.44 | 26.62 | 8 |
| 411412 | 90.81 | 10.61 | 8 |
| 422334 | 88.11 | 12.26 | 9 |
| 422533 | 70.45 | 28.03 | 10 |
| 423423 | 79.21 | 14.83 | 7 |
| 424425 | 83.00 | 15.77 | 7 |
| 424444 | 82.28 | 15.72 | 9 |
| 424524 | 63.44 | 21.08 | 8 |
| 523421 | 86.56 | 11.76 | 9 |
| 623322 | 72.50 | 29.28 | 8 |
| 623545 | 66.94 | 21.64 | 9 |
| 624422 | 75.00 | 16.20 | 7 |
| 624424 | 64.21 | 29.16 | 7 |
| 624525 | 65.56 | 20.42 | 9 |
| 624534 | 69.10 | 29.09 | 10 |
| 625655 | 43.31 | 27.89 | 8 |

**Table 6.4c: SG results for non-patients continued - gambles with a non-fatal treatment failure outcome**

| Health state being assessed | Treatment Failure | Mean | SD | n |
|---|---|---|---|---|
| 311222 | 525555 | 96.49 | 4.11 | 23 |
| 311222 | 623424 | 95.24 | 5.91 | 22 |
| 111311 | 322323 | 98.34 | 4.18 | 22 |
| 322323 | 623424 | 92.35 | 14.16 | 22 |
| 224244 | 623424 | 80.76 | 19.05 | 21 |
| 422413 | 623424 | 83.29 | 16.52 | 21 |
| 111312 | 521412 | 96.58 | 5.07 | 19 |
| 423122 | 623424 | 79.53 | 23.83 | 18 |

## Table 6.4d: SG results for patients

| Health State | 'Best' | | | | | | Min | Max |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | SE | Median | IQ range | n | Mean | Mean |
| 1) Treatment failure is death | | | | | | | | |
| 123143 | 78.54 | 26.55 | 4.00 | 92.50 | 65.0-99.50 | 44 | | 80.19 |
| 625655 | 63.30* | 30.30 | 4.79 | 60.00 | 35.0-95.00 | 40 | 59.45 | 53.65 |
| 2) Treatment failure is 625655 | | | | | | | | |
| 111212 | 78.00 | 25.85 | 7.46 | 90.00 | 45.0-99.0 | 12 | 73.42 | 83.92 |
| 111311 | 71.91 | 28.70 | 8.29 | 76.25 | 50.0-98.0 | 12 | 71.93 | 76.17 |
| 111312 | 87.04* | 10.89 | 3.28 | 85.00 | 80.0-99.5 | 11 | 84.75 | 90.70 |
| 124143 | 72.79 | 22.16 | 3.17 | 77.50 | 50.0-90.0 | 49 | 67.92 | 79.83 |
| 211111 | 82.14 | 24.11 | 6.45 | 95.50 | 50.0-99.5 | 14 | 72.08 | 84.58 |
| 224244 | 64.13* | 17.53 | 5.06 | 55.00 | 50.0-75.0 | 12 | 56.58 | 67.69 |
| 311222 | 73.04 | 23.18 | 6.20 | 75.0 | 50.0-95.0 | 14 | 63.23 | 82.64 |
| 313333 | 74.64 | 20.39 | 6.15 | 75.0 | 60.0-90.0 | 11 | 70.25 | 84.91 |
| 322323 | 73.68* | 18.58 | 4.97 | 70.0 | 60.0-90.0 | 14 | 67.64 | 79.36 |
| 422413 | 60.58* | 20.83 | 6.01 | 60.0 | 50.0-70.0 | 12 | 52.41 | 65.00 |
| 423122 | 70.71 | 20.68 | 5.97 | 78.75 | 45.0-90.0 | 12 | 67.53 | 80.33 |
| 524112 | 68.86* | 18.19 | 4.86 | 65.0 | 55.0-90.0 | 14 | 61.78 | 74.93 |
| 623424 | 56.99 | 20.30 | | | | 53 | 48.35 | 63.67 |

Note : * This indicates where the mean value exceeds the median

**Table 6.4e: Distribution of points of indifference higher than 99**

| Percentage chance of success | Frequency* |
|---|---|
| 99.50000 | 1 |
| 99.60000 | 3 |
| 99.66660 | 1 |
| 99.80000 | 8 |
| 99.86660 | 1 |
| 99.90000 | 12 |
| 99.98000 | 3 |
| 99.98990 | 1 |
| 99.99900 | 9 |
| 99.99980 | 2 |
| 99.99999 | 20 |
| Total | 82 |

* All non-patients, except 6 at 99.9999.

## Table 6.5a: Internal consistency amongst patients (n=55)

| Health State Comparison | | | Distance between states | VAS | | | Standard Gamble | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | > | = | < | > | = | < |
| 111111 | vs | 211111 | 1 | 11 | 1 | 2 | | | |
| 111111 | vs | 111212 | 2 | 13 | 0 | 0 | | | |
| 111111 | vs | 111311 | 2 | 12 | 0 | 2 | | | |
| 111111 | vs | 111312 | 3 | 9 | 0 | 3 | | | |
| 124143 | vs | 224244 | 3 | 11 | 1 | 1 | 6 | 3 | 3 |
| 211111 | vs | 311222 | 4 | 11 | 0 | 3 | 8 | 3 | 3 |
| 422413 | vs | 623424 | 5 | 10 | 0 | 3 | | | |
| 111111 | vs | 311222 | 5 | 12 | 0 | 3 | | | |
| 322323 | vs | 623424 | 6 | 12 | 0 | 2 | 8 | 1 | 5 |
| 111312 | vs | 313333 | 7 | 9 | 0 | 3 | 9 | 0 | 1 |
| 313333 | vs | 623424 | 7 | 13 | 0 | 0 | 7 | 1 | 3 |
| 423122 | vs | 623424 | 7 | 12 | 0 | 1 | 6 | 4 | 2 |
| 111311 | vs | 322323 | 7 | 12 | 0 | 2 | 8 | 1 | 3 |
| 111111 | vs | 423122 | 8 | 10 | 0 | 3 | | | |
| 111111 | vs | 523112 | 8 | 13 | 0 | 1 | | | |
| 111212 | vs | 422413 | 8 | 12 | 0 | 1 | 7 | 3 | 2 |
| 111111 | vs | 124143 | 9 | 61 | 0 | 7 | | | |
| 111111 | vs | 322323 | 9 | 12 | 0 | 2 | | | |
| 111111 | vs | 313333 | 10 | 9 | 0 | 4 | | | |
| 111212 | vs | 224244 | 10 | 12 | 0 | 1 | 7 | 4 | 1 |
| 311222 | vs | 623424 | 10 | 14 | 0 | 1 | 11 | 2 | 1 |
| 111111 | vs | 422413 | 11 | 12 | 0 | 1 | | | |
| 111111 | vs | 224244 | 12 | 13 | 0 | 0 | | | |
| 111312 | vs | 623424 | 12 | 10 | 0 | 2 | 10 | 0 | 1 |
| 111212 | vs | 623424 | 13 | 12 | 0 | 1 | 8 | 3 | 1 |
| 111311 | vs | 623424 | 13 | 14 | 0 | 0 | 7 | 0 | 5 |
| 211111 | vs | 623424 | 14 | 12 | 0 | 2 | 10 | 3 | 1 |
| 111111 | vs | 623424 | 15 | 50 | 0 | 5 | | | |
| | | | Distribution Total % | | | | | | |
| | | | | 413 | 2 | 53 | 112 | 28 | 31 |
| | | | | 88.2 | 0.5 | 11.3 | 65.5 | 16.4 | 18.1 |

## Table 6.5b Internal consistency amongst non-patients (n=110)

| Health state comparison | | | Distance between states | VAS | | | Standard Gamble | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | > | = | < | > | = | < |
| 111111 | vs | 211111 | 1 | 19 | 2 | 0 | | | |
| 111111 | vs | 111212 | 2 | 18 | 3 | 1 | | | |
| 111111 | vs | 111311 | 2 | 39 | 3 | 2 | | | |
| 111111 | vs | 111312 | 3 | 19 | 3 | 0 | | | |
| 124143 | vs | 224244 | 3 | 31 | 3 | 10 | 21 | 11 | 10 |
| 422413 | vs | 623424 | 5 | 21 | 0 | 1 | 16 | 2 | 2 |
| 111111 | vs | 311222 | 5 | 21 | 0 | 0 | | | |
| 111312 | vs | 521412 | 6 | 22 | 0 | 0 | 17 | 1 | 0 |
| 521412 | vs | 623424 | 6 | 21 | 0 | 1 | 15 | 1 | 2 |
| 623424 | vs | 624645 | 6 | 21 | 0 | 1 | 19 | 1 | 0 |
| 322323 | vs | 623424 | 6 | 21 | 1 | 0 | 16 | 3 | 3 |
| 111312 | vs | 313333 | 7 | 22 | 0 | 0 | 17 | 0 | 1 |
| 313333 | vs | 623424 | 7 | 20 | 0 | 2 | 13 | 4 | 1 |
| 423122 | vs | 623424 | 7 | 22 | 0 | 0 | 15 | 1 | 2 |
| 111111 | vs | 523111 | 7 | 22 | 0 | 0 | | | |
| 111111 | vs | 423122 | 8 | 22 | 0 | 0 | | | |
| 111111 | vs | 523112 | 8 | 44 | 0 | 0 | | | |
| 523111 | vs | 623424 | 8 | 21 | 0 | 1 | 19 | 1 | 0 |
| 111212 | vs | 422413 | 8 | 22 | 0 | 0 | 19 | 1 | 0 |
| 111111 | vs | 124143 | 9 | 129 | 1 | 0 | | | |
| 111111 | vs | 322323 | 9 | 22 | 0 | 0 | | | |
| 111111 | vs | 521412 | 9 | 22 | 0 | 0 | | | |
| 111111 | vs | 313333 | 10 | 22 | 0 | 0 | | | |
| 111212 | vs | 224244 | 10 | 22 | 0 | 0 | 19 | 0 | 1 |
| 311222 | vs | 623424 | 10 | 22 | 0 | 0 | 18 | 2 | 2 |
| 111111 | vs | 422413 | 11 | 22 | 0 | 0 | | | |
| 111111 | vs | 224244 | 12 | 44 | 0 | 0 | | | |
| 111312 | vs | 623424 | 12 | 22 | 0 | 0 | 18 | 0 | 0 |
| 124143 | vs | 525555 | 12 | 43 | 0 | 0 | 41 | 0 | 3 |
| 111212 | vs | 623424 | 13 | 22 | 0 | 0 | 18 | 0 | 0 |
| 111311 | vs | 623424 | 13 | 22 | 0 | 0 | 21 | 1 | 0 |
| 322323 | vs | 624645 | 13 | 22 | 0 | 0 | 19 | 1 | 0 |
| 211111 | vs | 623424 | 14 | 22 | 0 | 0 | 20 | 0 | 1 |
| 111111 | vs | 623424 | 15 | 87 | 0 | 0 | | | |
| 211111 | vs | 525555 | 19 | 22 | 0 | 0 | 20 | 0 | 1 |
| 111311 | vs | 624645 | 19 | 22 | 0 | 0 | 20 | 0 | 0 |
| 111111 | vs | 624645 | 21 | 22 | 0 | 0 | | | |
| 111111 | vs | 525555 | 20 | 21 | 0 | 0 | | | |
| 211111 | vs | 311222 | | 22 | 0 | 0 | 18 | 2 | 1 |
| 111311 | vs | 322323 | | 22 | 0 | 0 | 20 | 2 | 0 |
| Total n | | | | 1124 | 16 | 18 | 439 | 24 | 23 |
| % | | | | 97.1 | 1.4 | 1.6 | 90.3 | 4.9 | 4.8 |

Where: > Strictly consistent
= Equal response
< Strictly inconsistent

**Table 6.6: Summary of logical consistency**

| | Patients | | Non-patients | | Total | |
|---|---|---|---|---|---|---|
| Consistency[1] | n | % | n | % | n | % |
| VAS | | | | | | |
| > | 413 | 88.2 | 1124 | 97.1 | 1537 | 94.6 |
| = | 2 | 0.5 | 16 | 1.4 | 18 | 1.1 |
| < | 52 | 11.3 | 18 | 1.6 | 70 | 4.3 |
| SG | | | | | | |
| > | 112 | 65.5 | 439 | 90.3 | 551 | 83.9 |
| = | 28 | 16.4 | 24 | 4.9 | 52 | 7.9 |
| < | 31 | 18.1 | 23 | 4.8 | 54 | 8.2 |

1. where: > strictly consistent
= equal
< strictly inconsistent

**Table 6.7: Consistency by Distance between Health States**[1]

| Distance | | % Consistent | | Standard Gamble |
|---|---|---|---|---|
| | n | | n | % Consistent |
| 1 | 35 | 86 | - | - |
| 2 | 93 | 88 | - | - |
| 3 | 91 | 77 | 54 | 50 |
| 5 | 71 | 90 | - | - |
| 6 | 100 | 97 | 86 | 73 |
| 7 | 136 | 96 | 87 | 77 |
| 8 | 150 | 96 | 52 | 87 |
| 9 | 255 | 97 | - | - |
| 10 | 107 | 94 | 68 | 81 |
| 11 | 35 | 97 | - | - |
| 12 | 134 | 99 | 73 | 95 |
| 13 | 92 | 100 | 82 | 85 |
| 14 | 36 | 94 | 17 | 100 |
| 15 | 142 | 96 | - | - |
| 19 | 44 | 100 | 35 | 97 |
| 20 | 22 | 100 | - | - |
| 21 | 21 | 100 | - | - |

[1]: Where distance is the combined difference between the scale levels of the two states

**Table 6.8: Comparison of indirect and direct health state valuations**

| Assessed Health State | Worst Outcome | n | Value | | Indirect Value[1] | | Direct Value[2] | | Difference (Indirect-direct) |
|---|---|---|---|---|---|---|---|---|---|
| | | | mean | (SD) | mean | (SD) | mean | (SD) | |
| 311222 | 525555 | 22 | 96.67 | (4.11) | 98.29 | (2.42) | 92.18 | (8.90) | 6.11*** |
| 311222 | 623424 | 22 | 95.24 | (5.91) | 98.84 | (1.67) | 92.18 | (8.90) | 6.66*** |
| 111311 | 322323 | 22 | 98.69 | (2.20) | 9.90 | (0.42) | 98.34 | (4.18) | 1.56*** |
| 322323 | 623424 | 22 | 92.35 | (14.16) | 97.52 | (6.56) | 90.73 | (8.62) | 6.79*** |
| 224244 | 623424 | 20 | 80.76 | (19.05) | 92.88 | (10.48) | 82.65 | (16.25) | 10.23*** |
| 422413 | 623424 | 19 | 81.87 | (16.76) | 94.10 | (8.24) | 85.50 | (12.32) | 8.60** |
| 111312 | 521412 | 18 | 94.88 | (9.01) | 98.81 | (2.24) | 96.52 | (4.93) | 2.29** |
| 423122 | 623424 | 18 | 79.53 | (23.83) | 93.20 | (8.11) | 79.50 | (19.95) | 13.70** |
| 124143 | 625655 | 36 | 73.29 | (20.54) | 89.80 | (12.70) | 78.07 | (27.20) | 11.73*** |

1. Derived from the original value and transformed onto the scale defined by perfect health and death (see text)

2. Obtained from a single gamble with death as the worst state

**Table 6.9: Adjusted VAS data**

| Health State | Mean | SD | SE | Median | 1-2 range | n |
|---|---|---|---|---|---|---|
| 111111 | 100.00 | - | - | - | - | 155 |
| 111212 | 89.05 | 9.91 | 1.68 | 89 | 50-95 | 35 |
| 111311 | 87.72 | 12.38 | | | | 55 |
| 111312 | 90.84 | 7.74 | 1.39 | 90 | 88-95 | 31 |
| 124143 | 54.43 | 22.15 | 1.61 | 55 | 40-71 | 189 |
| 211111 | 86.97 | 12.97 | 2.22 | 89 | 82-95 | 34 |
| 222432 | 47.65 | 18.48 | 3.27 | 45 | 30-62 | 32 |
| 224244 | 42.24 | 28.53 | 3.51 | 40 | 31-63 | 66 |
| 311211 | 71.31 | 15.59 | 2.85 | 72 | 59-84 | 30 |
| 311222 | 67.16 | 15.03 | 2.58 | 70 | 55-77 | 34 |
| 313333 | 60.57 | 22.41 | 4.02 | 65 | 43-74 | 31 |
| 322323 | 61.86 | 19.38 | 3.37 | 65 | 49-75 | 33 |
| 323422 | 46.35 | 21.38 | 3.90 | 44 | 29-62 | 30 |
| 424413 | 54.54 | 24.18 | 4.09 | 58 | 37-74 | 35 |
| 422434 | 26.50 | 14.80 | 3.15 | 21 | 15-39 | 22 |
| 423122 | 54.76 | 17.67 | 3.18 | 53 | 40-71 | 31 |
| 521412 | 49.52 | 19.23 | 4.20 | 53 | 39-66 | 21 |
| 523111 | 62.62 | 19.85 | 4.23 | 63 | 53-80 | 22 |
| 525112 | 43.61 | 18.36 | | | | 55 |
| 525555 | 15.70 | 14.12 | 3.08 | 12 | 5-27 | 21 |
| 623424 | 32.52 | 18.96 | 1.64 | 30 | 20-45 | 133 |
| 624415 | 22.25 | 13.35 | 2.44 | 21 | 11-29 | 30 |
| 624645 | 11.37 | 14.71 | 3.21 | 11 | 4-20 | 21 |
| 625555 | 6.67 | 8.93 | 1.61 | 6 | 3-75 | 31 |
| 625655 | 15.33 | 24.44 | 3.27 | 11 | 5-28 | 56 |
| Death | | | | | | |

**Table 6.10 Comparison of Sheffield survey and the main MVH survey**

| Mean VAS score | SF-6D | | EQ-5D[1] | |
|---|---|---|---|---|
| | n | % | n | % |
| 0 - 19.9 | 5 | 9 | 4 | 10 |
| 20 - 39.9 | 14 | 25 | 17 | 41 |
| 40 - 59.9 | 18 | 32 | 9 | 21 |
| 60 - 79.9 | 12 | 21 | 7 | 17 |
| 80 - 99.9 | 8 | 14 | 5 | 12 |
| | 57 | | 42 | |

| Mean TTO/SG score | SF-6D SG | | EQ-5D TTO[1] | |
|---|---|---|---|---|
| | n | % | n | % |
| Below 0 | 0 | | 16 | 38 |
| 0 - 19.9 | 0 | | 5 | 12 |
| 20 - 39.9 | 0 | | 5 | 12 |
| 40 - 59.9 | 3 | 5 | 6 | 14 |
| 60 - 79.9 | 14 | 25 | 5 | 12 |
| 80 - 99.9 | 40 | 70 | 5 | 12 |
| | 57 | | 42 | |

1. Source: MVH, 1994

**Figure 6.1: Distributions of VAS values by health state**

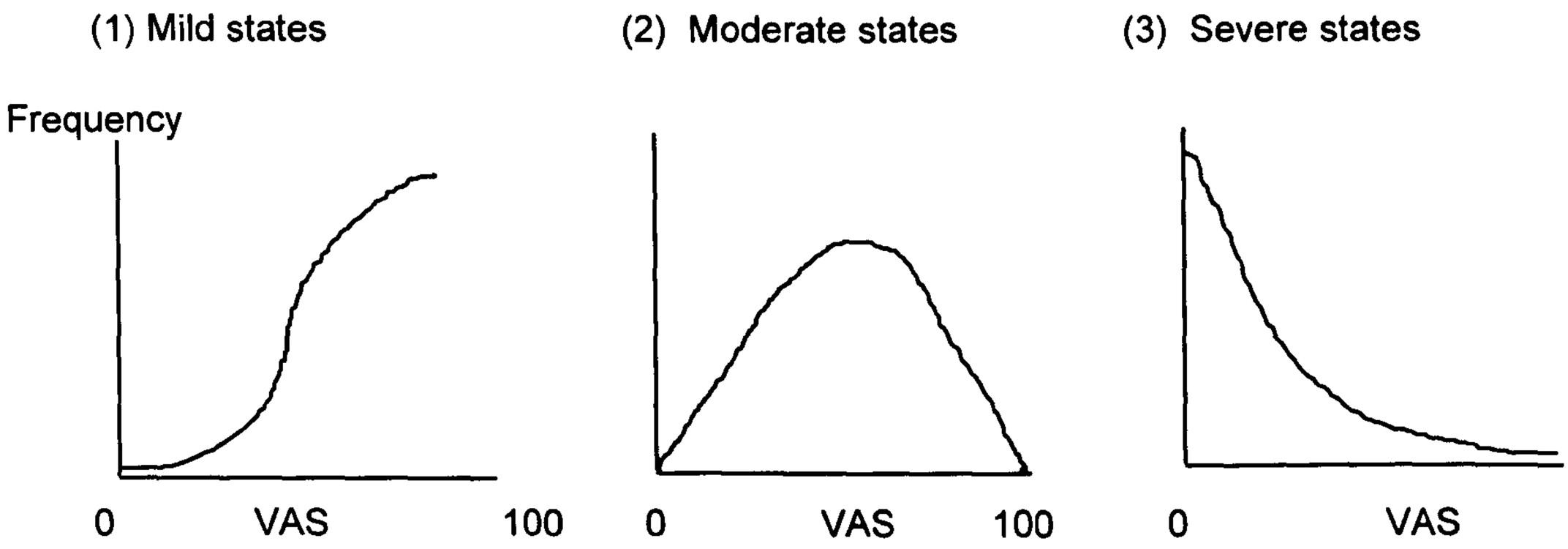


**Figure 6.2: Stem and Leaf Plot - VAS**

| Frequency | Stem & Leaf | | |
|---|---|---|---|
| 5.00 | Extremes | | (-111), (-78), (-56), (-44) |
| 1.00 | -2 | . | & |
| 1.00 | -2 | * | & |
| 3.00 | -1 | . | 8& |
| 2.00 | -1 | * | & |
| 8.00 | -0 | . | 556& |
| 3.00 | -0 | * | & |
| 14.00 | 0 | * | 000233& |
| 41.00 | 0 | . | 55555555555667788 |
| 45.00 | 1 | * | 000000000111111122344 |
| 47.00 | 1 | . | 5555555555556667788899 |
| 57.00 | 2 | * | 0000000001111111222223334 |
| 66.00 | 2 | . | 55555555555666667777777788899 |
| 63.00 | 3 | * | 000000000001111111122333333344 |
| 54.00 | 3 | . | 555555556666677777888889 |
| 69.00 | 4 | * | 00000000000001122222223444444444 |
| 57.00 | 4 | . | 55555555556666667777778899 |
| 62.00 | 5 | * | 00000000000000122222222233334 |
| 63.00 | 5 | . | 555555555555555556667777788888& |
| 61.00 | 6 | * | 00000000000011112222233333444 |
| 64.00 | 6 | . | 55555555556666666666777888888899 |
| 37.00 | 7 | * | 00000111222333344 |
| 51.00 | 7 | . | 55555556667777778888899 |
| 59.00 | 8 | * | 00000000011111222333333344444 |
| 49.00 | 8 | . | 5555555555566677788889999 |
| 47.00 | 9 | * | 000001111233444444444 |
| 25.00 | 9 | . | 55555666788 |
| 14.00 | 10 | * | 0000001 |
| .00 | 10 | . | |
| 1.00 | 11 | * | & |

Stem width: 10.00
Each leaf: 2 case (s)

& denotes fractional leaves.

**Figure 6.3: Stem and Leaf Plot - Standard Gamble**

| Frequency | Stem & Leaf | | |
|---|---|---|---|
| 58.00 | Extremes | | (0), (5), (10), (20), (25), (30), (35), (40) |
| 5.00 | Extremes | | (45) |
| 28.00 | 5 | * | 00000 |
| .00 | 5 | t | |
| 8.00 | 5 | f | 55 |
| 2.00 | 5 | s | & |
| .00 | 5 | . | |
| 21.00 | 6 | * | 0000 |
| 5.00 | 6 | t | 2 |
| 23.00 | 6 | f | 55555 |
| 8.00 | 6 | s | 77 |
| .00 | 6 | . | |
| 24.00 | 7 | * | 00000 |
| 16.00 | 7 | t | 222 |
| 40.00 | 7 | f | 55555555 |
| 15.00 | 7 | s | 777 |
| .00 | 7 | . | |
| 44.00 | 8 | * | 000000000 |
| 25.00 | 8 | t | 22222 |
| 58.00 | 8 | f | 555555555555 |
| 23.00 | 8 | s | 77777 |
| 7.00 | 8 | . | 9& |
| 78.00 | 9 | * | 0000000000000000 |
| 29.00 | 9 | t | 222223 |
| 74.00 | 9 | f | 555555555555555& |
| 170.00 | 9 | s | 66666666666666667777777777777777 |
| 228.00 | 9 | . | 88888888888888899999999999999999999999999999 |
| 4.00 | 10 | * | 0 |

Stem width: 10.00
Each leaf: 5 case (s)

& denotes fractional leaves.

# Chapter 7

## Estimating the relationship between Standard Gamble and Visual Analogue Scale Valuations

The administration of SG alongside VAS to value the same sample of health states provides an opportunity to compare them. This is of considerable theoretical and practical interest.

Past comparisons of these elicitation techniques have tended to yield the same basic result, that SG values significantly exceed those of VAS (Torrance, 1976; Bombadier et al., 1982; Llewellyn-Thomas et al., 1984; Read et al., 1984; Bass et al., 1994). One study found the reverse, but this was not significant in the statistical sense (Hornberger et al., 1992), and the MVH pilot survey found a cross-over at around 0.8, with milder health states having lower SG values than the adjusted VAS ratings (Dolan et al., 1995a). There have been a number of attempts to explain these differences between SG and VAS (Bombadier et al., 1982; Torrance et al., 1992; Loomes, 1993; Dolan and Sutton, 1997), and these have drawn on important theoretical developments in the literature on decision-making. However, the empirical work has been limited by one or more of the following aspects: only one theory being used to explain the differences; using aggregate level analysis when the theories are concerned with individual behaviour; and/or they have been based on insufficient numbers of observations. By overcoming these drawbacks, this chapter seeks to provide further insight into an important theoretical debate.

There is an important practical objective to this research. The VAS technique is easier and therefore cheaper to administer than the SG and has been shown to achieve better levels of completion, consistency and reliability in this and other surveys (e.g. Dolan et al., 1996). This was the reason for Torrance and his co-workers (1992) choosing to elicit preferences for the HUIs using VAS, and to estimate a power function between VAS and SG in order to transform the VAS values into SG values (see Chapter 3). The SG technique also presents ethical problems with particular patient groups, since it presents potentially upsetting scenarios involving death (Drummond and Davies, 1991). As

reported in the previous chapter, attempts to overcome this problem by changing the worst reference state leads to significant violations of EUT. There would be considerable benefits for research, therefore, if it were possible to estimate a relationship between VAS and SG.

The chapter begins with a discussion of alternative theoretical explanations for the relationship between VAS and SG. This is followed by sections on the methods of modelling the relationship, a presentation of results using the Sheffield valuation survey, and a discussion of the implication of the results for this study and other applications.

## 7.1 Theoretical explanations

### 7.1.1 Relative risk attitude

An explanation for the differences in the values obtained by these elicitation techniques is that VAS generates a value under certainty whilst SG generates utility value under uncertainty (Torrance, 1976; Gafni and Birch; 1993; Bowe, 1995). According to this explanation, SG and VAS will only be the same for individuals who are risk-neutral. The relationship between SG and VAS depends on a person's attitude to risk in ways depicted on Figure 7.1 for three relative risk attitudes. A risk-averse person has a concave utility function, indicating he/she would prefer a certain health state with a value x to an expected equivalent value x calculated by summing two or more health state values by their probability. The risk seeker would have a convex utility function, indicating the opposite. For a risk-neutral person, the two curves will be the same.

It is usual to assume a rational individual would have a constant attitude to risk (Currim and Sarin, 1984), which is the assumption underlying the risk-adjusted QALY model (Chapter 2.4.1). This risk-adjusted QALY model implies the following function:

$$U(x) = [v(x)]^r \qquad (1)$$

The parameter r is a person's relative risk attitude, where $r>1$ implies risk seeking, $r<1$ implies risk aversion and $r=1$ risk neutrality. U(x) is a von-Neumann Morgenstern or

SG utility function and v(x) a value function. Dyer and Sarin (1982) describe the value function in the following way: "*We are left with introspection based on the assumption that strength of preference is a primitive concept. Several assessment procedures accept this viewpoint, including direct rating, the use of the direct ordered matrix and exchange questions (see Fishburn, 1967 for a review), but none of these approaches can be verified by actually observing choices by the decision-maker'' (P877) (emphasis added).* VAS is a version of direct rating, therefore equation (1) can be regarded as representing the relationship between SG and VAS[1].

For respondents with relative risk aversion (RRA), this theory predicts a relationship between SG and VAS represented by a concave curve where all points lie above the $45^{o}$ time and bow out in a northward direction in the manner shown on Figure 7.1.

Doubts have been raised about this theory. Most empirical work testing this relationship has used group mean health state values, but evidence at an individual level suggests models based on power functions may not fit the data as well as linear models (Dolan and Sutton, 1995/6). Loomes and his colleagues (1994) found that even at the aggregate level, the parameter estimates were not robust[2]. It is therefore important to consider other explanations for the relationship between VAS and SG.

### 7.1.2 Gambling effect

Bombardier et al. (1982) explained the pattern in terms of "a general aversion to gambling with one's health, a 'gambling aversion' which must be distinguished from the 'risk aversion' familiar to students of decision analysis" (P. 152; also quoted in Loomes (1993)). The existence of this general aversion to gambling in SG utilities has been acknowledged by some health economists (Gafni, 1994). This was the model proposed by Morrison (1994) to explain the inconsistencies between direct and indirect

---

[1] V (X) has been presented as a TTO value (e.g. Johannesson, 1994), but it is questionable whether TTO can be used to derive a measurable value function since Dyer and Sarin (1982) argue it cannot be verified by observing choices.

[2] Loomes has also proposed Regret theory as an alternative explanation for curved relationships between TTO and VAS as well as SG and VAS (Loomes, 1993). However, there was no means of testing this theory with the data collected in the survey.

SG valuations (see Chapter 6). It has been argued by Richardson (1994) that this gambling effect is not allowed for in EUT and offers a different or complementary explanation to RRA.

### 7.1.3 Framing effect

This explanation for the relationship between VAS and SG stems from the different reference points implied by the elicitation tasks (Loomes et al., 1994; Dolan and Sutton, 1997). In a SG question, a respondent is asked to imagine that he/she is in a chronic health state, certain to last ten years. he/she is then asked to consider a risky treatment option. The chronic state of each SG question therefore becomes the reference state. Loomes et al. (1994) and Dolan and Sutton (1997) have argued that in VAS, the respondent would take full health as their reference point "*on the entirely reasonable grounds that she is currently in normal health and has not been asked to suppose otherwise*" (Loomes et al., 1994 p11).

According to Kahnerman and Tvesky's (1979) Prospect Theory an individual considers outcomes as either gains or losses relative to their perceived reference point. Kahnerman and Tvesky have proposed a value function concave in gains but convex and steeper in losses (Figure 7.2). This would imply risk aversion over gains and risk seeking over losses, rather than a constant attitude to risk. Loomes et al. (1994) have shown how this function, combined with the different reference points of VAS and SG, can generate the non-linear relationship between SG and VAS shown in Figure 7.3[3].

---

[3] In their example, a value function is assumed to weight losses three times as much as corresponding gains (the exact weighting does not alter the general result). They considered five equidistant health states, where J > S > R > N > K. On the basis that y (J) = 0, the reference point of VAS, all states are seen as losses and have the values of -210, -330, -405 and -450 respectively. Rescaling these values so that J is 100 and K is zero produces scores of 53 for S, 27 for R and 10 for N. The reference point of the SG questions can be S, R or N, with J being a gain and K a loss. For an SG question involving S as the chronic state, y (s) becomes zero, y (I) = +70 and y (k) = -405. Transforming y (s) onto a scale where J equals 1.0 and K is zero results in the indifference point for S (i.e. where it is most difficult to choose between state S and a risky treatment involving J and K) being achieved when P is 0.85. The corresponding values of P for SG involving R and N are 0.75 and 0.61 respectively. These values have been reproduced in the following table:

| | Reference point | | | |
|---|---|---|---|---|
| Health state | y (j) = 0 | y (s) = 0 | y (r) = 0 | y (y) = 0 |

As noted by Loomes et al. (1994), it will be difficult to distinguish between this 'reference point plus value function' (RPVF) explanation and RRA from just SG and VAS data, since the only difference is that under RRA it is conventionally assumed that risk attitude is constant, whereas it can vary under the RPVF explanation.

## 7.2 Methods

### 7.2.1 Specification

There are ten specifications for the relationship between VAS and SG examined in this study, and these are presented on Table 7.1.

The general aversion to gambling can be represented by a linear equation with an intercept and a slope defined by VAS. More complex linear models include quadratic and cubic VAS terms for non-linearities in the relationship.

The RRA explanation is represented by a standard power function:

$$U(X) = b_1 [V(X)]^{b2} \qquad (1)$$

A similar pattern can be generated by an alternative power function, originally proposed by Torrance (1976) to describe the relationship between VAS and TTO but since re-

| | Score | VAS[1] | Score | P[1] | Score | P | Score | P |
|---|---|---|---|---|---|---|---|---|
| J | 0 | 100 | 70 | 1.00 | 110 | 1.00 | 135 | 1.00 |
| S | -210 | 53 | 0 | 0.85 | 70 | | 110 | |
| R | -330 | 27 | -210 | | 0 | 0.75 | 70 | |
| N | -405 | 10 | -330 | | -210 | | 0 | 00.61 |
| K | -450 | 0 | -405 | 0 | -330 | 0 | -210 | 0 |

Source: Loomes et al., 1994

1. Calculated by setting J to 100/1.0 and K to zero.

Plotting these P values against the rescaled VAS scores generates a relationship similar to the one predicted by RRA (Figure 7.3).

specified for relating VAS to SG (Torrance et al., 1982; Loomes 1993; Dolan and Sutton, 1995)[4]:

$$U(X) = 1 - b_1 [(1 - V(X)]^{b2} \qquad (2)$$

where $b_1$ is a constant and $b_2$ the power term. In Torrance's work the constant $b_1$ has been restricted to unity, but it would be preferable not to limit its value (Dolan and Sutton, 1997). In the specification examined in this research, risk aversion is represented by $b^2 > 1$ and risk seeking by $b^2 < 1$.

Outside of the field of health, Currim and Sarin (1984) have recommended the following exponential relationship:

$$U(X) = (1 - e^{-cv)})/(1-e^{-c}) \qquad (3)$$

The parameter c is a constant and reflects the person's RRA. The person is relatively risk averse when $c < 0$; risk neutral when $c = 0$ and risk seeking when $c<0$.

Ten model specifications have been examined in this chapter: three versions of the 'linear' function (models 1a - 1c), three power functions (models 2a - 2c), and three versions of Torrance's power function (models 3a to 3c). Each version places different restrictions on the parameter values. These models allow for the possibility of gambling aversion combining with RRA or RPVF by including a constant term in some of the power models (i.e. 2c and 3c). A final model is the exponential function suggested by Currim and Sarrin (1984).

---

[4] The original basis for the relationship was not RRA. The initial formulation was $VAS = 1 - (1 - TTO)^{b1}$, and Torrance claimed support for this from the psychometric evidence on the relationship between VAS and magnitude estimation. In his earlier study, Torrance found SG to be equivalent to TTO, and hence used the same formulation for the relationship between VAS and SG. He now accepts the RRA explanation (e.g. Torrance et al., 1995).

The impact of the background characteristics of respondents on the relationship will be considered in the individual level analysis in terms of differences by age, sex and self-rated health.

### 7.2.2 The data sets

These models have been estimated on two data sets. The first was the aggregate level data set, where the regression analyses have been undertaken on mean health state values. The second was the individual observations. The VAS and SG data sets are those from the non-patient sample described in Chapter 6. They exclude certain respondents for gross inconsistencies and the VAS data has been transformed so that zero is death and one is full health to allow aggregation across individuals.

To estimate the models, it has been necessary to transform to VAS and SG scores from a zero to 100 scale onto a zero to 1.0 scale. It has also been necessary to replace 28 negative VAS values (i.e. states worse than death) with small positive values (i.e. 0.01) and 13 values of one or more by 0.9999. Background characteristics are entered as a dummy variable for age (under/over 40), sex, the presence of perceived chronic ill health and a five category self-rated health item ( i.e. item 1 of the SF-36)[5].

### 7.2.3 Estimation and testing

For these analyses, Ordinary Least Squares (OLS) regression has been used. It has been suggested that for the VAS and SG data the appropriate technique would be a Tobit model because the values have a limited range (i.e. 0 to 1.0) (Dolan and Sutton, 1995b). This technique is normally used on a censored or truncated data set when only a part of some larger distribution is available for analysis, such as the examples of a poverty line in an income distribution or the demand for tickets to a football game when the stadium is at full capacity, and are typically associated with a large number of observations at one or other end of the distribution (Greene, 1993). This is not the case for either VAS or SG values. There are very few observations at either the floor (i.e. 0) or ceiling (i.e. 1.0), though there is a clustering of SG values towards the ceiling. Therefore, Ordinary Least Squares regression should be adequate.

---

[5] 'In general would you say your health is: Excellent, very good, good, fair, or poor ?'

SG values are, however, positively skewed and this may lead to violations of the assumptions of classical regression of constant variance and normality in the error terms. A way to overcome this problem is to use a transformation function which maps the unit interval (0, 1) onto an infinite line (-α.,α). A logit transformation of the SG values will be examined to see if this improves the models.

The goodness of fit of the nested models is usually compared using the F-test (Greene, 1993), and this has been undertaken in the comparisons of versions 2 and 3 of the linear and power functions. However, the first versions of these functions do not have a constant term. This alters the meaning of the R-squared since the mean error is no longer zero and it is generally not recommended to use it in such circumstances (Stewart and Wallis, 1981). The method for testing the impact of individual coefficients in such circumstances is the conventional t-test. It is also not possible to compare formally the goodness of fit of the linear and power models.

Models have been tested for the normality errors, heterogeneity and general specification. Normality of residuals was formally tested by the Kolmogorov-Smirnov test. Heteroscedasticity has been tested by regressing the square of the residuals against the predicted value and performing an F-test of significance. A general test of specification has been undertaken by Ramsey's RESET test, where the square of the predicted values are included in a second run of the model (Ramsey, 1969). An F-test was undertaken to assess the significance of any improvement in R-squared for the nested models with constant terms.

### 7.2.4 Analysis plan

The overall aim of the analysis was to select the best model for the aggregate and individual data sets using the conventional criteria of parsimony, goodness of fit (where it can be compared) and the diagnostic test results. The process is to select the best of each of the four types of model (i.e. linear, power, Torrance's power and exponential) before undertaking a comparison of the different functional forms. A comparison is also

undertaken with models estimated for the logit of SG. The predicted SG from the parameter estimates of the best models are then plotted against VAS.

## 7.3 Results

The Spearman rank correlation between the 58 mean SG and VAS health state values was 0.89 and the product moment correlation was 0.87. This level of agreement hides an important pattern in the disparities. For all health states, the SG mean and median values were above the VAS ratings (Table 7.2). Furthermore, these differences appear to be related to the health state. The pattern can be seen most clearly in the plots of SG and VAS mean (Appendix Plot A6.1) and median (Appendix Plot A6.2) values. All points are above the $45^{o}$ line (i.e. where VAS = SG) and there is some evidence of a bowing outwards of the relationship in a northward direction.

The rank correlation between SG and VAS of the 961 individual observations was 0.57. In the vast majority of cases SG exceeds VAS. A plot of SG against VAS reveals considerably more dispersion than at the aggregate level and a concentration of points near the ceiling of the SG scale (Appendix Plot A6.3 ). There was also some suggestion of bowing outwards in a northward direction.

### 7.3.1 Aggregate level

The results of modelling mean SG and VAS health state values are presented on Tables 7.3 and 7.4. All ten models were significant, and able to explain more than 80% of the variation in SG in six cases. The quadratic model was a significant improvement over the simple linear specification in terms of adjusted R-squared and passed the specification test. The cubic term was not a significant improvement in terms fit over the quadratic specification, and hence the latter has been selected from the three linear models (i.e. 1a - 1c). In common with the other models, however, it suffered from significant heterogeneity, with residuals declining against predicted SG.

The most general of the power models (i.e. model 2c) provided the best fit of the three, though it also suffered from heterogeneity (Table 7.4). Out of the three Torrance-based

power models, the most general version does not provide a significantly better fit, but in contrast to the other two specifications passed the specification test.

The exponential model was only able to explain 15% of the variation, and failed all diagnostic tests.

The selected versions of the linear, power, and Torrance's power models are therefore models 1b, 2c and 3c. These models are not nested, and therefore cannot be formally compared, though they achieved very similar levels of explanatory power (i.e. 82% - 83%) and the same result against the three diagnostic tests. They all suffered from heterogeneity. A re-run of these models using the logit of SG resulted in the quadratic and power models passing the heterogeneity test (Table 7.5).

Plots of the three selected versions of the functions for mapping VAS into SG values are shown in the Appendix (Plots A6.4-6.6). They all have a very similar shape. The intercepts are predicted to be 0.51, 0.23 and 0.49 for the quadratic, the power and Torrance's power function respectively. The curves are bowed outwards in a north-easterly direction with a declining gradient. The graphs indicate a cross-over of the $45^{o}$ line for the quadratic and Torrance's power functions at 0.95.

### 7.3.2 Individual level

The results of the individual level modelling are shown on Tables 7.6 - 7.9. All except two of the models achieved significance and explained between 25-28% of the variation. The exceptions were the first of the Torrance's power models, whose fit was worse than the mean value on its own, and the exponential model with an R-squared of 0.09.

The cubic function generated the best fit of the three linear models by significantly improving the R-squared (Table 7.6). However, all the linear models failed the normality and heterogeneity diagnostic tests. The cubic model only failed the specification test at the 5% level whereas the other two failed it at the 0.1% level. The residuals again showed a tendency to decline against predicted SG values.

Between the power models, model two was a better fit than model 1, but this was not significant. Model two was also superior in terms of the RESET specification test. All models failed the tests of heterogeneity and non-normality, though models 2 and 3 passed the RESET test. Model 3 is preferred to model 1 and 2, because the constant term was found to be significantly different from zero (Table 7.9).

The most general version of the Torrance's power function was also found to have a significantly better fit than the other two, though it also fails the tests of normality and heterogeneity. The exponential model explained rather less of the variation and failed all diagnostic tests.

The selected versions of the three types of model are therefore 1c, 2c and 3c. The level of explanatory power achieved was similar, but the cubic linear model failed the specification test. All models suffered significant levels of non-normality in their residuals and heterogeneity. A further run of these models using a logit of SG did not overcome these problems (Table 7.9).

The chosen functions for mapping VAS to SG values have been plotted (Appendix Plots A6.7-9). The cubic and the most general versions of the power function and Torrance's power function have a non-zero intercept. They also share the basic concave shape of the functions at the aggregate level, with the exception of the cubic function which has a slight upturn at the end. These functions predict that when VAS equals 1.0 the SG will have values of 1.0 for the cubic model, 0.98 for the power function and 0.93 Torrance's. The cubic function therefore crosses the $45^{o}$ line at 1.0, and the others are below this point. Torrance's power function has the lowest cross-over point.

The background variables were not found to improve significantly the performance of the three selected models in terms of fit, and the diagnostic tests (Table 7.9).

## 7.4 Discussion

The results of the valuation survey presented confirmed earlier findings that SG and VAS health state valuations are significantly correlated, but SG values usually exceed VAS. The precise form of this relationship has been explored by estimating and testing a range of models on aggregate, that is average health state values and individual level data.

### 7.4.1 Aggregate level

At the aggregate level, the preferred linear, power and Torrance's power models were able to explain 81-83% of the variation, but they each suffered from a significant degree of heterogeneity. There was little to choose between these models, though the quadratic and Torrance's power model predict a cross-over at 0.95 and this is not compatible with actual data nor the theoretical explanations considered here. For this reason, the power function would be preferred. The heterogeneity in the model was resolved by transforming the skewed SG values using a logit function, but this cannot be easily related to any of the theoretical explanations.

### 7.4.2 Individual analysis

At the individual level, the explanatory power of the models was less impressive. The preferred versions of the linear, power and Torrance's power specifications were only able to achieve an R-squared of 28-29%, and all suffered from non-normal residuals and heterosedasticity. These problems were not resolved by the logit transformation of the SG values. The cubic linear model failed the general test of specification, whereas the two power models passed this test which suggests they are better specified. This finding differs from another recent study which has presented detailed modelling work at the individual level. Dolan and Sutton (1995) found the linear models to achieve a better fit than the power functions.

The final choice is therefore between two power specifications. One method of choosing between them is to consider the credibility of the predictions. When VAS is unity, Torrance's power model (3c) predicts SG to be 0.93 and the power model (2c) predicts

0.98. Such crossovers occurred in models estimated by Dolan and Sutton (1997). Inspection of the plots shows there are very few observations where VAS ratings exceed SG utilities. The cross-over is also inconsistent with the RRA and RPVF explanations. Therefore there is a case for choosing 2c over 3c on the grounds that it is more consistent with the data.

The very poor predictive ability of the exponential model was also found by Loomes et al. (1994). These findings contrast with Currim and Sarin (1984), who found this model minimised the Sum of Squared Errors for 40 out of 43 individuals compared to a linear function, although they did not present the goodness of fit of the models. One explanation for this discrepancy could be the differences in the subject matter. In their study, Currim and Sarin asked about preferences between jobs, where the students responding to the questions may have had a more consistent and well-defined set of preferences.

### 7.4.3 Theoretical implications

These results have implications for the different theoretical explanations. All models predict a positive non-zero intercept (where they are free to do so). The best versions of the linear, power and Torrance's power specifications have a value between 0.31 and 0.54 for the constant term. These findings are consistent with a general aversion to gambling.

The power function and Torrance's power function had significant power terms which were consistent with relative risk aversion. These findings are consistent with the RRA and the RPVF explanations, and account for the concavity of the relationship. RRA and RPVF could be regarded as competing explanations, though as Dyer and Sarin (1982) have suggested they may also be complementary: the individual has a constant RRA and the value function could be 'S' shaped around an individual's reference point. Furthermore, given the poor fit of these models at the individual level and the extent of the heterogeneity and non-normal in the residuals, there are likely to be other unmeasured sources of variation between individuals. It has been suggested by Read et al. (1984) and Revicki (1992), for example, that there might be other psychological

explanations. However, to disentangle these competing explanations it would have been necessary to conduct interviews with respondents in order to understand the cognitive processes involved in undertaking the two elicitation tasks.

### 7.4.4 Practical implications

An important reason for conducting this analyses has been the potential practical benefits of being able to use the VAS instead of SG in terms of greater ease of completion and acceptability to respondents, more reliable data, and less evidence of respondent confusion. This was the reason for Torrance and his colleagues using VAS to value the different versions of the HUI (see Chapter 3).

The poor explanatory power of the models, less than 30%, and the evidence of non-normality and heterogeneity, suggest VAS is not able to predict SG at the individual level. The results would seem to be more promising, however, at the aggregate level. The high level of explanatory power at this level was comparable to results achieved in two earlier studies. Bombadier et al. (1982) estimated a linear model with an $R^2$ of 0.76 and subsequently Loomes (1993) was able to fit Torrance's original function to the same data with an $R^2$ of 0.80.

The relationship between VAS and SG health state values, even at this aggregate level, were not the same in these studies. The parameters had the same sign, but their size was not the same as those found in this research. In a linear regression, Bombardier and colleagues estimated a significant constant of 0.32 and a slope coefficient of 0.88 compared to 0.60 and 0.46 respectively found in this study. The power coefficient found in this study for Torrance's specification was 2.16 compared with 4.89 in the original work by Torrance et al., 1982. In a more recent study by Dolan and Sutton (1997) using the same versions of the VAS and SG questionnaires, they found the parameter estimates were different in sign as well as magnitude.

The review in Chapter 5 found considerable scepticism among economists and others regarding the cardinal properties of the VAS technique as means of eliciting strength of preference. VAS valuations are subject to a 'spreading effect', whereby respondents

seek to use the entire length of a scale regardless of the severity of the condition being valued. There is also a context effect, whereby valuations depend on the severity of the other states being valued. Interviews with respondents have found that the meaning given to the VAS exercise has differed from the strength of preference interpretation, and included notions of chronology and degrees of physical fitness.

These explanations would suggest that little significance can be attached to the differences between VAS scores, and hence its relationship to SG is likely to vary between studies. For these reasons, Loomes et al. (1994) concluded that *"even if there appears to be a systematic general relationship between VAS scores and SG utilities, there seem no straightforward way of converting one into the other which is stable across procedures and contexts"*. It would therefore be inappropriate to use VAS scores to predict SG when it is possible to obtain SG values directly.

## 7.5 Conclusion

There is little evidence for the theoretical explanations of the relationship between VAS and SG at the individual level. A better relationship was found between VAS and SG mean health state values, but the parameter in the models was different from those found in other studies. There is no theoretical or empirical support for mapping VAS scores into SG utilities at the aggregate or individual level. This would raise doubts about the validity of the algorithms for estimating SG utilities for the HUI Marks II and III, since these were based on such a transformation. The implications for the research presented in this thesis are that the modelling of health state values presented in the next chapter must be undertaken with actual SG data rather than values extrapolated from the VAS data, despite the latter being better in terms of completeness, more reliability and more consistency.

**Table 7.1: Model specifications**

| | |
|---|---|
| 1a Simple linear | $U = b_0 + b_1 V$ |
| 1b Quadratic | $U = b_0 + b_1 V + b_2 V^2$ |
| 1c Cubic | $U = b_0 + b_1 V + b_2 V^2 + b_3 V^3$ |
| 2a Power (1) | $U = Vb^2$ |
| 2b Power (2) | $U = b_1 Vb^2$ |
| 2c Power (3) | $U = b_0 + b_1 Vb^2$ |
| 3a Tpower (1) | $U = 1 - (1 - V)^{b2}$ |
| 3b Tpower (2) | $U = 1 - b_1 (1 - V)^{b2}$ |
| 3c Tpower (3) | $U = b_0 - b_1 (1 - V)^{b2}$ |
| 4 Exponential | $U = (1 - e^{-cv(x))})/(1 - e^{-c})$ |

**Table 7.2: Comparison of SG and VAS average values - common states**

| Health State | Standard Gamble | | Visual analogue scale | | Difference (SG - VAS) | |
|---|---|---|---|---|---|---|
| | mean | median | mean | median | mean | median |
| 111212 | 96.22 | 99.25 | 92.58 | 94.59 | 3.64 | 4.66 |
| 111311 | 98.76 | 99.00 | 88.15 | 92.00 | 10.61 | 7.00 |
| 111312 | 95.69 | 97.50 | 90.84 | 93.00 | 4.85 | 4.50 |
| 124143 | 88.13 | 95.00 | 51.07 | 53.00 | 37.06 | 42.00 |
| 211111 | 96.61 | 99.00 | 88.70 | 89.00 | 7.91 | 10.00 |
| 222432 | 91.98 | 95.50 | 47.65 | 45.00 | 44.33 | 50.50 |
| 224244 | 88.14 | 92.75 | 40.40 | 39.00 | 47.74 | 53.75 |
| 311211 | 97.14 | 98.00 | 71.31 | 72.00 | 25.83 | 26.00 |
| 311222 | 92.18 | 96.00 | 67.64 | 71.00 | 24.54 | 25.00 |
| 313333 | 81.46 | 95.00 | 58.37 | 64.00 | 23.09 | 31.00 |
| 322323 | 90.73 | 90.00 | 65.04 | 65.00 | 25.69 | 25.00 |
| 323422 | 90.98 | 95.50 | 46.35 | 44.00 | 44.63 | 51.50 |
| 422413 | 83.05 | 86.25 | 58.00 | 60.00 | 25.05 | 26.25 |
| 422434 | 88.68 | 94.00 | 26.50 | 21.00 | 62.18 | 73.00 |
| 423122 | 79.50 | 85.00 | 52.60 | 53.00 | 26.90 | 32.00 |
| 521412 | 82.61 | 90.00 | 49.50 | 53.00 | 33.11 | 37.00 |
| 523111 | 88.81 | 94.00 | 62.61 | 63.00 | 26.20 | 31.00 |
| 525112 | 90.36 | 96.00 | 40.78 | 40.00 | 49.58 | 56.00 |
| 525555 | 50.46 | 50.00 | 15.70 | 12.00 | 34.76 | 38.00 |
| 623424 | 77.58 | 84.00 | 31.46 | 30.00 | 46.12 | 54.00 |
| 624415 | 82.97 | 85.00 | 22.25 | 21.00 | 60.72 | 64.00 |
| 624645 | 62.17 | 70.00 | 11.37 | 11.00 | 50.80 | 59.00 |
| 625555 | 54.34 | 60.00 | 6.66 | 6.00 | 47.68 | 54.00 |

**Table 7.3: Linear Regression models with mean health state values (coefficients and SEs in parenthesis) [1]**

| Independent Variable | (1a) | (1b) | (1c) |
|---|---|---|---|
| Constant | 0.60 (0.02)*** | 0.51 (0.07)*** | 0.47 (0.03)*** |
| V | 0.46 (0.04)*** | 0.96 (0.16)*** | 1.42 (0.28)*** |
| V2 | - | -0.52 (.11)*** | -1.69 (0.67)* |
| V3 | - | - | 0.812 (0.46) |
| df | 54 | 53 | 52 |
| Adjusted $R^2$ | .75 | .82[2] | .82 |
| Normality | NS | NS | NS |
| Homogeneity | *** | ** | * |
| Specification | *** | NS | NS |

1. The multivariate models have been run on 0 to 1.0 scales (i.e. not 0 to 100) and asterisks indicate a significant improvement over previous (nested) model

**Table 7.4: Power and exponential models with mean health state values**

| Independent Variable | (2a) | (2b) | (2c) | (3a) | (3b) | (3c) | (4) |
|---|---|---|---|---|---|---|---|
| Constant | Zero | Zero | 0.23 (0.17)*** | Unity | Unity | 0.95 (.02)*** | - |
| V | Unity | 0.99 (0.14)*** | 0.77 (0.16)*** | Unity | 0.47 (0.03)*** | 0.46 (0.03)*** | - |
| Ln (V) | 0.22 (0.01)*** | 0.22 (0.02)*** | 0.30 (0.10)** | - | - | | - |
| Ln (1 - V) | - | - | - | 4.41(0.36)*** | 1.72 (0.14)*** | 2.41 (0.39)*** | - |
| C | - | - | - | - | - | - | 4.89 (0.36)*** |
| df | 55 | 53 | 54 | 55 | 54 | 53 | 55 |
| Adjusted $R^2$ | 0.79 | $0.80^{+}$ | $0.83^{+}$ | 0.06 | $0.81^{+}$ | 0.82 | 0.15 |
| Normality | NS | NS | NS | | NS | NS | ** |
| Homogeneity | * | * | ** | | ** | *** | *** |
| Specification | NS | NS | NS | *** | * | NS | *** |

**Table 7.5: Models of VAS against the logit of SG mean health state values**

| Independent Variable | (1b) | (2c) | (3c) |
|---|---|---|---|
| Constant | 0.13 (0.21) | 0.09 | 4.07*** |
| V | 3.13 (0.95)** | 3.72*** | 4.00*** |
| $V^2$ | 0.62 (0.94) | - | - |
| Ln (V) | - | 1.06*** | - |
| Ln (1 - V) | - | - | 0.83*** |
| df | 53 | 53 | - |
| Adjusted $R^2$ | 0.80 | 0.81 | 0.80 |
| Normality | NS | NS | NS |
| Homogeneity | NS | NS | *** |
| Specification | NS | NS | * |

**Table 7.6: Linear Regression models with individual values**

| Independent Variable | (1a) | (1b) | (1c) |
|---|---|---|---|
| Constant | 0.67 (0.12)*** | 0.58 (0.02)*** | 0.52 (0.02)*** |
| V | 0.36(0.02)*** | 0.81(0.08)*** | 1.55 (0.18)*** |
| $V^2$ | - | - 0.45 (0.08)*** | -2.28 (0.42)*** |
| $V^3$ | - | - | 1.21 (0.27)*** |
| df | 961 | 960 | 959 |
| Adjusted $R^2$ | 0.25 | 0.28$^+$ | 0.29$^+$ |
| Normality | *** | *** | *** |
| Homogeneity | *** | *** | *** |
| Specification | *** | *** | * |

+ indicates a significant improvement over the previous model ($\alpha < 0.05$)

**Table 7.7: Power and exponential regression models with individual values**

| Independent Variable | (2a) | (2b) | (2c) | (3a) | (3b) | (3c) | (4) |
|---|---|---|---|---|---|---|---|
| Constant | Zero | Zero | 0.31 (0.14)* | Unity | Unity | 0.93 (0.01)*** | - |
| V | Unity | 0.97 (0.01)*** | 67 (0.13)*** | Unity | .40 (0.02)*** | 0.39 (0.02)*** | - |
| Ln (V) | 0.18 (0.01)*** | 0.15 (.01)*** | 0.26 (0.08)*** | - | - | - | - |
| Ln (1 - V) | - | - | - | 7.25 (0.32)*** | 1.60 (0.11)*** | 3.14 (0.34)*** | - |
| Exponent | - | - | - | - | - | - | 7.21 (0.27)* |
| df | 962 | 961 | 960 | 962 | 961 | 960 | 962 |
| Adjusted $R^2$ | 0.28 | 0.29$^{+}$ | 0.29 | -.15$^{++}$ | 0.27 | 0.29$^{+*}$ | .09 |
| Normality | *** | *** | *** | *** | *** | *** | *** |
| Homogeneity | *** | *** | *** | *** | *** | *** | *** |
| Specification | *** | NS | NS | *** | *** | NS | *** |

+ Indicates a significant improvement over the previous model ($\alpha$ <0.05)

++ NB The negative value indicates that this model had a worse fit than the mean

**Table 7.8: Selected regression models of VAS against logit of SG with Individual values**

| Independent Variable | (1c) | (2c) | (3c) |
|---|---|---|---|
| Constant | 0.13 (0.28) | 0.44 (0.30) | 5.02 (0.31)*** |
| V | 9.12 (2.31)*** | 4.32 (0.28)*** | 4.40 (0.29)*** |
| $V^2$ | -12.53 (5.33)* | - | - |
| $V^3$ | 8.59 (3.51)* | - | - |
| Ln (V) | - | 0.95 (0.16) | - |
| Ln (1 - V) | - | - | 0.87 (0.15) |
| Exponent | - | - | - |
| df | 953 | 954 | 954 |
| Adjusted $R^2$ | 0.25 | 0.25 | 0.25 |
| Normality | | *** | *** |
| Homogeneity | | ** | ** |
| Specification | NS | NS | NS |

**Table 7.9: Effect of background variables**

| Independent Variable | Model (1c) |
|---|---|
| Constant | 0.54 (0.03)*** |
| V | 1.56 (0.18)*** |
| $V^2$ | -2.30 (0.42)*** |
| $V^3$ | 1.23 (0.27)*** |
| Ln (V) | - |
| Ln (1 - V) | - |
| Age (over 45) | 0.01 (0.01) |
| Chronic | -0.00 (0.01) |
| Sex | 0.01 (.01) |
| General Health | 0.00 (0.01) |
| df | 955 |
| Adjusted $R^2$ | 0.29 |
| Normality | *** |
| Homogeneity | *** |
| Specification | *** |

**Figure 7.1: The relationship between utility and value under three types of relative risk attitude**

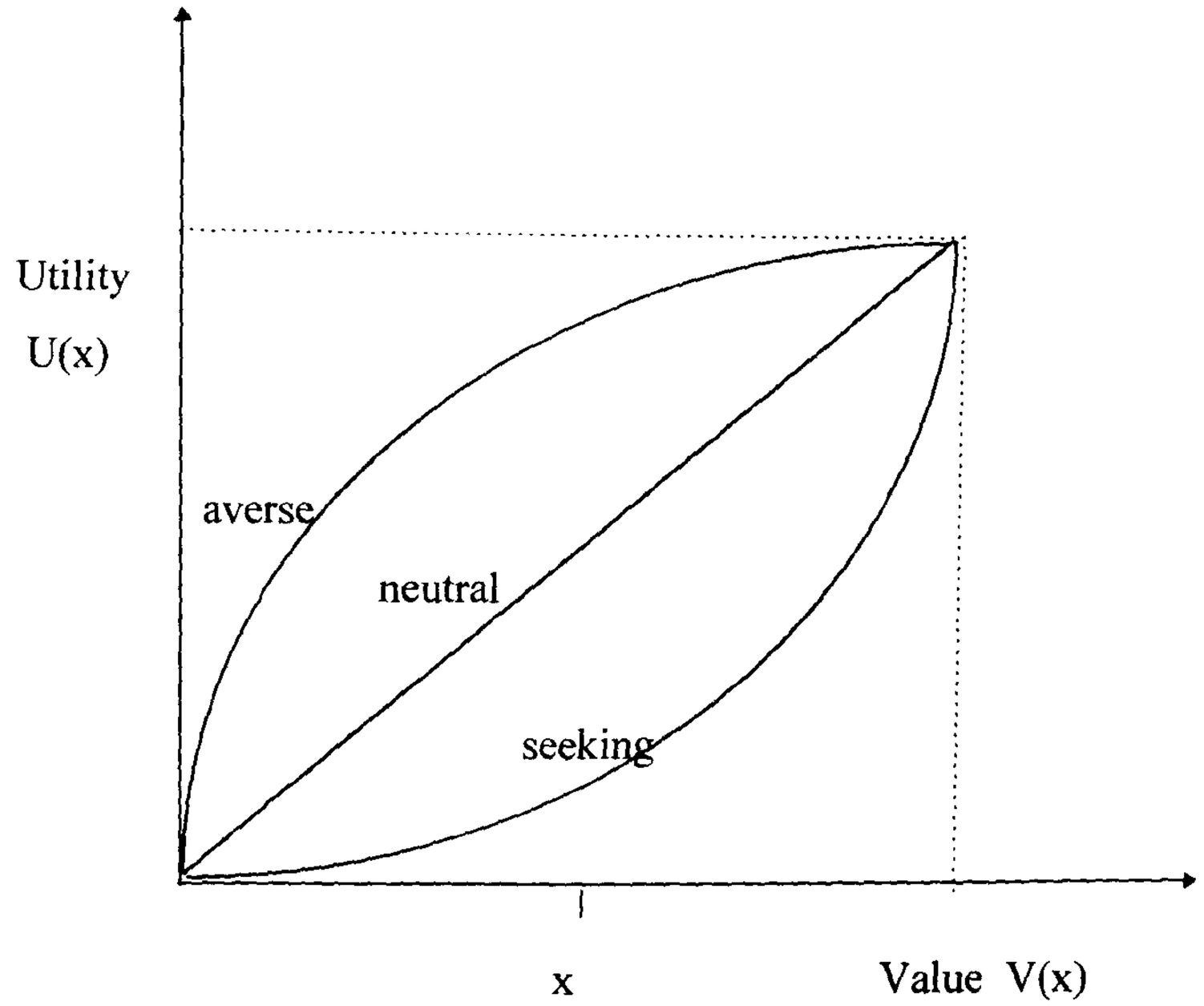


**Figure 7.2: The Kahneman and Tversky value function**

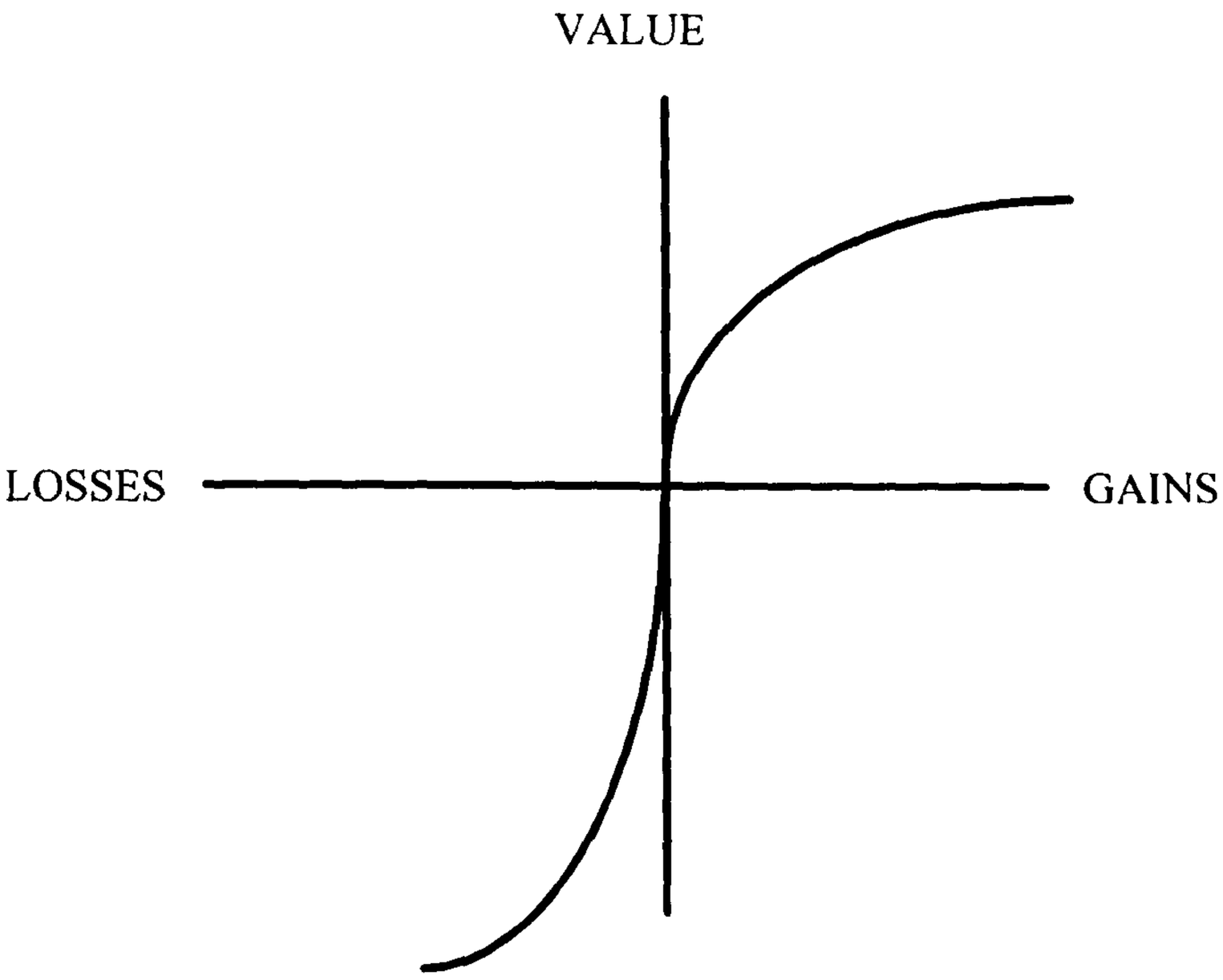
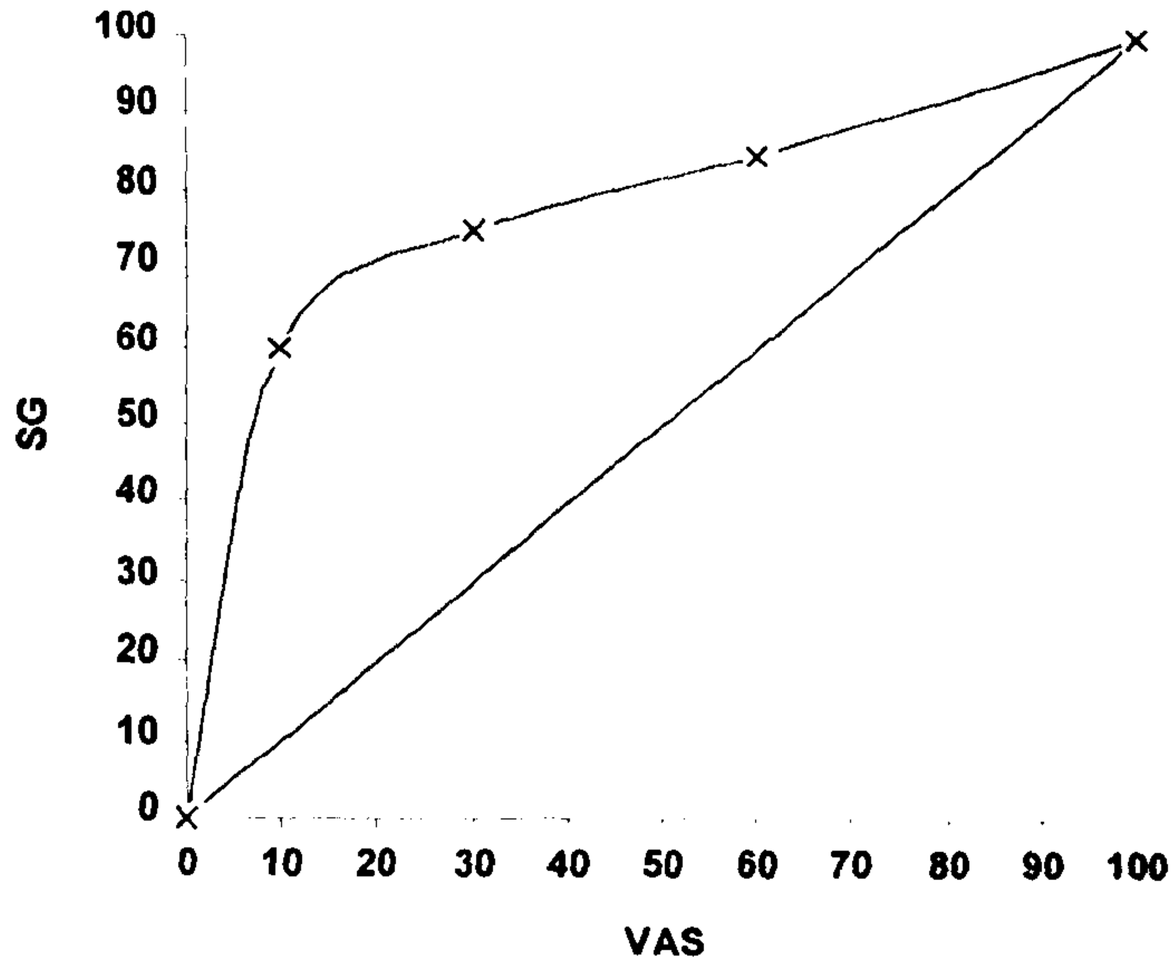



**Figure 7.3: Relationship between SG and VAS predicted by RPVF theory**



# Chapter 8

# Modelling Values for the SF-6D

The purpose of the econometric modelling presented in this chapter is to predict VAS and SG values for all health states defined by the SF-6D from the sample of 57 states valued in the survey. The chapter begins by examining the methodological problems of model specification and estimation and the chosen solutions. The results of the modelling are then reported and discussed.

## 8.1 Methods

### 8.1.1 Dependent variables

The preparation of the VAS and SG data sets for analysis has been described elsewhere (Chapter 6.8). To recap, there were 165 respondents who provided 1582 VAS ratings and 1567 SG values for 57 health states (excluding state 111111 and death). To permit comparisons between respondents, VAS data have been adjusted by transforming the results onto a scale of 1.0 for state 111111 and zero for death (equation 6.1). After this adjustment, and the exclusions for major inconsistencies, there were a total of 1357 VAS observations by 155 respondents. There were also some exclusions from the SG data set for major inconsistencies. The main exclusion was those gambles with a non-fatal outcome failure because these were found to produce values onto the full health to death scale that was significantly different from those obtained directly. All patient respondents, who mainly undertook non-fatal gambles, have therefore been excluded. This left 1037 SG observations from 106 respondents.

The distributions of the VAS ratings and SG values were found to be skewed. There are arguments for using either the mean or median as the measure of central tendency, therefore both have been modelled (see arguments in Chapter 6.9). The median models have used the median value for each health state as the dependent variable. The mean models have been undertaken at the individual level, where each valuation is regarded

as a separate observation, rather than using the mean value for each health state, since this makes better use of the data and greatly increases the number of degrees of freedom available for the analysis (from 56 to over 1000).

Individual level analysis also allows an adjustment to be made for the effect of the respondent on the health state values. Respondents did not value the same set of states and although a balanced design was used in selecting health states for each individual (see Chapter 6), differences between health state values may be partly due to differences in the preferences of the respondents who valued them rather than the attributes of those states. Disentangling the respondent effect is a complex task, and is most comprehensively undertaken at the individual level.

Skewness in the dependent variable may result in heterogeneity in the error term of a regression model. It is therefore important to consider transformations (MacCullogh and Nelder, 1983). Another problem arises from the fact that all the mean and median values and 99% of individual values, lie between zero and one. This may lead to violations in the assumptions of classical regression of constant variance as well as normality in the error terms. A method of avoiding this problem is to use a transformation function on the dependent variable which maps the unit interval (0, 1) onto an infinite line (- ∞, ∞). Logit transformations of the dependent variable achieve this and will be investigated in order to examine whether this improves the models[1].

### 8.1.2 Specification of models

The simplest functional form is the additive model, where the levels of each dimension are entered as dummy variables i.e.

[1] Abdalla and Russell (1995) have also examined two complementary log - log functions in their modelling of the EQ-5D using MVH data.

$$y_i = \alpha + \sum_{k,j}^{k=6,\, j=n} \beta_{kj}\, x_{kj} + e_i \qquad (1)$$

This model has a constant term $\alpha$ and a set of dummy variables $x_{kj}$ for each level j of dimension k of the SF-6D. Level one of dimension 3 (i.e. social functioning), for example, is denoted by variable $x_{31}$. For any given health state, $x_{kj}$ will be defined as follows:

$x_{kj} = 1$ if, for this state, dimension k is at level j

$x_{kj} = 0$ if, for this state, dimension k is not at level j.

In all, there are 23 of these terms, with level 1 acting as the baseline for each dimension. The value of health state 111111 is by implication the intercept term $\alpha$. The value of all other states is derived by summing the coefficients of the 'on' dummies.

This specification restricts the model to an additive form, but it imposes no restrictions on the size of the intervals between the levels of the dimensions. For example, it does not enforce an equal interval scale. Earlier work with the Euroqol found the assumption of equal interval to be invalid for certain dimensions (van Hout and McDonell, 1991, and MVH, 1994). Such an assumption is likely to be even more dubious for the larger dimension scales of SF-6D. Furthermore, the additive model proposed here does not impose ordinality on the levels. This allows the respondents to confirm or otherwise the judgements of the developers of the SF-6D where there was ambiguity in the ordering of levels (e.g. limitations in walking 100 yards Vs. limitations in bathing and dressing) and to assess respondent understanding where ordering is unambiguous (e.g. very severe bodily pain should not be ranked as better than severe pain).

There could be interactions between health dimensions. Torrance et al. (1992) have suggested *".... that the additional disutility added by a particular deficit is greater if it is the first and only deficit and less if it is the last of two or more deficits."* Alternatively, for some states an interaction may increase the deficit over and above the

sum of the two parts. However, the estimation of all possible interaction terms would have required a substantially larger proportion of the 9000 potential health states of the SF-6D to be valued. It is not be possible to include all 236 first order interactions in the median models, since these have just 23 degrees of freedom. Models using individual data have around 1000 degrees of freedom, and therefore could examine the first order interactions. However, with so many first order interactions, there is a risk of finding significant interactions due to the play of chance. The modelling has therefore been restricted to those interactions between significant main effects.

To extend the modelling to higher order interactions another specification has been used. Variables have been defined by the extreme levels of each dimension of the SF-6D. These are denoted by dummies for the number of times a health state contains dimensions at the extreme ends of the scale (MVH, 1995). The least severe has been defined as the first level on each dimension, and the most severe end of each dimension has been defined to include the worst two levels[2]. The extreme variables are denoted by $E_{1m}$ and $E_{2m}$, the least and most severe respectively, where m = 1, 2, .., 6 and describes the number of times the least or most severe levels appear in a state. Thus, for example, $E_{1m}$ assumes a value of 1 if the number of level ones in a health state is equal to m, and zero if not. In all, there will be 12 of these dummy variables. The model therefore becomes:

$$yi = \alpha + \sum_{k,j}^{k=6,\,j=n} \beta_k x_{kj} + \sum_m \gamma_{1m} E_{1m} + \sum_m \gamma_{2m} E_{2m} + e_i \qquad (2)$$

The variables $E_{1m}$ and $E_{2m}$ allow for non-linearities in the relationships between levels of the dimensions at the extreme ends of their scales. Combination of levels can have a greater or lesser effect than the sum of the parts. These extreme dummy variables are tested on both median health state and individual level models.

---

[2] It has been appropriate to vary the number of dimensions of the most severe as follows: PF-5 & 6; RF-2; SF-5; Pain - 5 & 6; MH-4 & 5; V-4 & 5)

### 8.1.3 Estimation techniques

For the aggregate models, the median health state values have been obtained from health states with different numbers of observations. This may give rise to a form of heteroscedasticity, since the variance of the error term will depend on the number of observations (Stewart and Wallis, 1982). All else being equal, the variance will be inversely dependent on the number of observations (n). The appropriate estimation technique is therefore Weighted Least Squares (WLS), which assumes the error terms are independently normally distributed random variables with a standard deviation weighted by n. This has been estimated using the Statistical Package for Social Sciences (SPSS).

For the individual level model, inter-respondent variation must be taken into consideration in the estimation. Preferences are likely to vary between individuals and this could lead to bias in the estimates generated by Ordinary Least Squares (OLS). The usual approach is to assume that differences in "tastes" can be accounted for by a set of dummy variables for identifiable characteristics of respondents such as age, sex and health status (e.g. Bates, 1988). However, respondents values or preferences for health states may vary in ways which are not explained by these background characteristics. Many differences in "taste", for example, will depend on attitudinal variables unobserved in the study. Health state values are therefore likely to be clustered by respondent, and this leads to the data having a multi-level structure as shown in Figure 1. There are two sources of variation in the data sets: within respondent (level 1) and between respondent (level 2). In this study, individuals have valued different sets of health states, and hence there is a risk of confounding between the values assigned to health states and the values of respondents, leading to bias the estimated coefficients. Furthermore, the variation between observations from different respondents is likely to exceed the variation in values from the same respondent. The error terms are not independent, and hence an assumption of OLS is violated (Greene, 1993).

One solution to the problem of biased coefficients is the inclusion of a dummy variable for the fixed effect of each respondent which results in the following model:

$$y_i = \alpha_0 + \alpha_r + \sum \beta_{kj} x_{kj} + e_i \qquad (3)$$

where $\alpha_0$ is an overall constant, and $\alpha_r$ denotes a specific constant attributable to respondent r. This allows the health state valuations of each respondent to deviate from the population average by some constant amount. For example, a respondent may tend to place all health states further up a VAS rating scale than the average respondent. In SG valuations, someone with a larger aversion to gambling (see Chapter 7 for explanation) may tend to value health states closer to 100. This is known as a 'fixed-effects' model (Goldstein, 1993) or the 'least squares dummy variable model' (Greene, 1995).

Fixed-effects models such as equation (3) can be estimated in a single stage OLS procedure. However, it is more convenient, particularly when examining large numbers of interactions, to break it down into two stages. In the first stage, a model is estimated with just the respondent specific constant terms, and the residuals are saved. In the second stage, the residuals are used as the dependent variable. This two-stage procedure has been used here.

The fixed-effects model specified in equation (3) is limited because it only allows the value of the constant term to vary. Respondents are likely also to differ in the weight they give to dimensions, and to the intervals between levels of dimensions. These variations could be examined in a fixed-effects model by having interaction terms between respondent and each dimension level, but this would create too many terms to be estimated with these data sets. Therefore, a more sophisticated technique must be used which allows the coefficients to vary randomly between respondents. This is known as a random effects model (Greene, 1993), or a multi-level model (Goldstein, 1995).

A random effects method can be used in the first instance as an alternative way of allowing for variations in the constant term. This specification assumes each respondent

has his/her own intercept term which is randomly distributed about the population mean:

$$\delta_r = \alpha + U_r$$

where $\alpha$ is the mean constant term and $U_r$ the variation between individuals, with a mean of zero, variance of $\sigma_u^2$ and covariance $(U_i, U_r) = 0$. Incorporating this into equation (1) yield:

$$yi = \alpha + \sum_{k,j} \beta_k x_{kj} + (U_r + ei) \qquad (4)$$

The fixed component is the same as for equation (1), but there is now a random component indicated in brackets and this contains two levels: the variation within respondent ($e_i$) and the variation between respondent ($u_r$). This specification provides a more efficient way of incorporating between respondent variation than equation (3), since it uses fewer degrees of freedom. However, it requires the assumption that respondents were randomly selected from their populations.

Variations between respondents can take more complex forms. The coefficient of each explanatory variable can be assumed to vary randomly between respondents about the population average as follows:

$$\beta_{kjr} = \beta_{kj} + U_{kjr}$$

where $U_{kjr} \sim N(0, \sigma_{ukj}^2)$

Feeding into equation (4):

$$y_i = \alpha + \sum_{k,j} \beta_{kj} x_{kj} + (U_r + U_{11r} + \ldots + U_{kjr} + \ldots + e_i) \qquad (5)$$

The random component of the equation (in brackets) is now considerably enlarged and incorporates a separate error term for each level of each dimension. This complex error structure has been estimated using the statistical package MLn (Woodhouse et al., 1995). However, the size of such a model for the SF-6D would be too large (even for

MLn). It was therefore necessary to estimate the random effects separately for each dimension.

### 8.1.4 Testing

The goodness of fit of the OLS models has been examined in terms of the adjusted $R^2$, and comparisons made between nested models using the F-test. MLn uses a maximum likelihood method of estimation, and therefore nested models have been compared in terms of the log likelihood ratio. However, it was not possible to compare the goodness of fit of OLS models with those estimated by MLn.

The conformity of the models to the assumptions underlying the estimation techniques has been examined by a series of diagnostic tests. Normality has been formally tested by the Kolmogorov-Smirnov test. The significance of heteroscedasticity was examined by regressing the square of the residuals on the predicted values and performing an F-test of significance of the model. A Ramsey RESET test has been used for assessing misspecification.

The robustness of the parameter estimates to changes in the sample has been examined by running the models on two samples of the individual data and comparing the results by performing a Chow test (Gujarati, 1993). A further assessment of robustness has been undertaken by re-running the model after excluding outliers from the tail of distribution of SG values.

### 8.1.5 Analysis plan

The plan of analysis is summarised on Figure 2. The aim is to find the best models for predicting VAS and SG values for health states defined by the SF-6D. Goodness of fit and conformity with the assumptions of the estimation technique are used for testing the specification of the different models. However, it is also important to be consistent with logical ordering of the scales of the SF-6D, where there is an unambiguous ranking (e.g. severe pain Vs very severe pain). Parsimony is also important, since the final algorithm must be transparent and readily understandable to other potential users. Any added

complexity, such as transformations or the addition of interactions, must be justified by significant and substantial improvements to the model. The final selection of models will therefore be based on the multiple criteria of goodness of fit, consistency with the SF-6D and parsimony.

Four models will be selected: VAS median, VAS mean (based on individual data), SG median, and SG mean. The robustness of these chosen models will be assessed. Finally, any dimension levels found to be inconsistent will be merged to create models that are consistent with the SF-6D.

## 8.2 Results

### 8.2.1 Visual analogue scale

Median values

The results of regressing median health state values against the SF-6D using Weighted Least Squares (WLS) are presented on Table 8.1. The model has an adjusted $R^2$ of 0.96 with 12 of the 23 SF-6D variables significant at the 5% level, and it passed the tests of normality in its residuals, homogeneity and misspecification. A logit transformation of the VAS values did not improve the fit of the model and neither did the inclusion of the extreme variable interaction terms (Appendix 8, Table A8.1)[3]. The simple additive function with main effects was found to be the best model for this data set.

The estimated coefficients of the SF-6D dummy variables have the expected negative sign in all cases, except three non-significant positive values. The rankings of the coefficients are consistent with the ordinality of the dimensions of the SF-6D for 19 out of the 23 adjacent pairs of levels. The exceptions being PH4 to PH5 (i.e. between levels four and five of the physical functioning dimension), Pain 2 to Pain 3, Pain 5 to Pain 6 and M4 to M5. The latter two were large and significant. To remove these inconsistencies it has been necessary to merge levels of the physical, pain, mental and vitality dimensions and create new variables PH45, Pain 23, Pain 456, M345 and V23

[3] The letter A before the number of a table indicates it is located in the appendix of this chapter.

(Table 8.1). Enforcing consistency reduced the number of SF-6D variables to 15, but did not significantly lower the explanatory power of the model (F-test at 5% level).

Individual values

The fixed-effects adjustment for the respondent effect improved the fit of the model from an adjusted $R^2$ of 0.55 to 0.68, and increased the number of significant SF-6D variables from 14 to 17 (Table 8.2). The model passed all diagnostic tests. A logit transformation of the VAS values and the inclusion of extreme variables did not improve the fit of the model (Appendix, Table A8.2a). There were 48 first interaction terms from all significant and consistent main effects, and these were entered into a stepwise regression with the significant and consistent main effects. This did not improve the adjusted $R^2$, and only two of the interaction terms were found to be significant and these were strongly collinear with their main effects. The simple additive function has therefore been selected on the grounds of parsimony, since no improvement in goodness of fit has been achieved from these changes.

Running the model on two sub-samples of the data SF-6D resulted in parameter estimates of a similar magnitude and the same sign (Appendix 8, Table A8.2b). The Chow test did not find any significant differences between the models. The exclusion of outliers, amounting to 2½% of observations, also had little effect on the parameter values (Appendix 8, Table A8.2c).

The estimated parameter coefficients of the model were negative in all cases except one. The rankings of the coefficients were consistent for 19 adjacent pairs of dimension levels, the exceptions being P2 to P3, P5 to P6, M4 to M5, and V3 to V4. The exclusion of these inconsistencies through the merging of adjacent pairs and the exclusion of S2 for being positive, resulted in a reduced version of the model (Table 8.2), but it did not significantly reduce the explanatory power of the model.

### 8.2.2 Standard gamble

Median values

The SF-6D model estimated by WLS had an adjusted $R^2$ of 0.90, and seven significant dimension level coefficients (Table 8.3). There was evidence of a significant degree of misspecification, though the model passed the diagnostic tests for homogeneity and normality. The logit transformation and the addition of extreme variable dummies for possible interactions did not improve the fit of the model and none of the extreme variable coefficients was found to be significant (Appendix 8, Table A8.3). The simple additive function achieves as good or better fit than the more complex models, but there was evidence of misspecification and a low number of significant coefficients.

All coefficients on the dimension levels have been estimated to be negative except for three non-significant positive estimates. There were four inconsistent adjacent pairs (S4 to S5, Pain 3 to Pain 4, M3 to M4, and V4 to V5) and an inconsistency between PH4 and PH6. Ten variables were merged or excluded in order to achieve consistency, but this did not significantly reduce the explanatory power of the model, and it resulted in the model passing the general specification test.

Individual values

The fixed-effects adjustment significantly improved the fit of this model from 0.324 to 0.492 (Table 8.4a)[4] and resulted in seven significant dimension levels out of the 23. The residuals were found to deviate from normality, and there was evidence of significant heterogeneity, though the model passed the general test of specification.

A logit transformation of the SG values resulted in a model with a lower explanatory power but there is some evidence of an improvement in the fit of the model since the number of significant SF-6D variables has increased from seven to ten and the number

---

[4] The adoption of the mid-point between the lower and highest chances of success as a proxy for indifference could be wrong for cases where the respondent places a tick against 100% and a cross against 99%. This extrapolation assumes the respondent would have taken a risk (i.e. a chance of success of 0.995), when he/she could be indicating an unwillingness to take any risk. The model was therefore re-estimated assuming a value of 1.0 for all observations with a mid-point of 0.995 (n = 41), and this was found to have no impact on the size or significance of the coefficients (Table A8.4e).

of inconsistencies has been reduced (Table 8.4b). However, there was no improvement in terms of non-normality and heterogeneity, and the model failed the general specification test. Given the equivocal nature of this evidence, and the importance of parsimony, it was decided to select the model using the untransformed dependent variable.

The addition of extreme variable terms for interactions did not improve the explanatory power of the model, though it reduced the number of significant coefficients and increased the number of inconsistencies (Appendix 8, Table A8.4a). Entering first order interactions also did not improve the fit of the model. Just five of the first order interaction terms were retained by the stepwise procedure, and four main effects were lost. Again, the significant interactions were strongly collinear with the main effects.

The models run on two sub-samples of the data were similar and there was no significant difference between them (Appendix 8, Table A8.4b). The exclusion of outliers (SG values less than 0.5 and 0.25) also did not change the sign of the coefficients, nor substantially alter their magnitude (Appendix 8, Table A8.4c). The main exceptions were those with the largest coefficients in the original, notably Pain 5, Pain 6 and Mental health 5, whose size was reduced[5].

Only two of the SF-6D variables had positive coefficients and these were not significantly different from zero. There were five inconsistencies between dimension levels of the SF-6D: PH4 to PH5 (i.e. 'limitations in 100 yards' was valued less than 'limitations in ½ a mile'), S2 and S3, Pain 2 and Pain 3, M3 and M4, and V3 and V4 (though most of these inconsistencies involved differences of less than 0.01). The merging of these dimension levels reduced the size of the model to 15 terms but did not reduce its explanatory power (Tables 8.4a).

[5] The model was also run on the SG data set with the mid-point value of 0.995 replaced by 1.0 in order to answer a concern raised in chapter 6. Again, there is no substansive impact on the coefficients ( A9.4d).

### 8.2.3 Impact of introducing a multi-level error structure

The detailed results from running the random effect or multi-level models are presented in tables in Appendix 8 ( Tables A8.5a & b and A8.6a & b), and summarised here.

The VAS and SG models with random components for the respondent's own constant term had similar, though not identical, coefficients to those estimated using fixed-effects adjustments. The differences were small and could be the result of using MLE compared with OLS rather than the impact of the random effects adjustments. The addition of dimension levels into the random part of the models improved the fit of the model (i.e. a significant improvement in the log-likelihood ratio), but did not substantively change the coefficients on the dimension levels. For the VAS model, four out of the six dimensions had significant variance or covariance terms (extra to the constant variance), but there were no comprehensible patterns to these terms. The SG models had more significant terms in their random components than the VAS model, and there was a positive association between the size of the variance terms and dimension level. This would suggest that between respondent variation in SG valuations increased with the severity of illness.

The more complex error structures improved the efficiency of the models, but did not substantively alter the coefficients of the dimension levels in the models and therefore added complexity was not justified.

### 8.2.4 Comparisons between actual and estimated values

Mean and median values have been estimated for 23 common health states using the consistent versions of the selected models described above. For the median models, this is a straightforward prediction using the estimated coefficients presented in Tables 8.1 and 8.3. For the individual level models, an adjustment was made to the constant term since the coefficients presented in Tables 8.2 and 8.4 predict the residuals of the respondent models (estimated in the first stage). In order to predict VAS or SG values,

it was necessary to add a mean respondent effect given by their mean values. The estimated and actual health state values are presented in Table 8.5a & b and 8.6a & b.

The standard errors of the estimated median VAS health state values were between 0.02 and 0.03 (Table 8.5a & b). The largest absolute difference between the actual and estimated health state values was 0.1 (for health states 322323 and 422434). In the majority of cases, the differences were 0.05 or less and there was no discernible pattern to them.

For the estimated mean VAS health state values, standard errors were between 0.01 and 0.02. The largest absolute difference between actual and estimated was 0.12 (health state 422434). Most differences were below 0.06, and again there was no discernible pattern.

The standard errors around estimated median SG health state values were between 0.01 and 0.04. The absolute differences between actual and estimated values were 0.05 or less for all health states except health state 422434 that had the largest difference of 0.07.

The standard errors around the mean SG health state estimates were 0.01 to 0.02 and differed from actual values by 0.06 or less. For both SG models, there was no pattern to the differences between the estimated and actual health state values.

## 8.3 Discussion

### 8.3.1 The models

The execution of the analysis plan has yielded a set of models for estimating median and mean VAS and SG values for health states defined by the SF-6D. The selected models are all additive linear, since entering interaction terms did not improve their fit and resulted in inconsistencies. For three of the models, no improvement was achieved from taking the logit of the dependent variable. The exception was an improvement in the SG

individual data model from an increase in the number of significant coefficients, but it failed the specification test. It was therefore decided not to use the logit transformation.

The fixed-effects models were found to be unbiased but inefficient compared to the more complex random effects models. There was evidence to suggest between respondent variation in SG valuations increased with the severity of illness. This would seem to reflect a heterogeneity in the variance of SG values observed in Chapter 6. However, the small improvements in efficiency were not sufficient to justify a multi-level model. Furthermore, these models assume the respondents are selected randomly, which is incorrect for these data sets. The fixed-effects adjustment therefore is preferred.

Many of the coefficients for the dimension levels were not statistically significant at the 5% level. For the median models this was partly due to having only 33 degrees of freedom. This was not a problem for the individual models. The SG individual model had evidence of heterogeneity which would have raised the size of the standard errors, and this partly explains the low number of significant coefficients.

The inconsistent rankings of the dimension levels may have arisen from the collinearity between the independent variables and this is shown on the correlation matrix of the SF-6D dimensions (Table 8.7). As discussed in Chapter 5, it was not possible to use a factorial design in the choice of health states for the survey. In real life, the different dimensions of health do not occur independently. The solution was to merge levels in order to achieve consistency. This did not significantly reduce the goodness of fit of the models.

The consistent versions of the VAS models and the SG median model passed the three diagnostic tests. Only the SG individual model failed the tests of normality and heterogeneity. Transforming the dependent variable did not resolve these problems, and neither did the addition of interaction terms. There is little else which can be done, since it is not possible to transform the independent variables given their categorical nature. The same problem was encountered in the modelling undertaken to value the EQ-5D by the MVH (Dolan, 1995).

The models were found to be robust, as indicated by the split sample test. A further test of removing outlier values had the effect of reducing the magnitude of the severe levels of mental health and pain for SG mean model. This is an unsurprising result since health states containing these levels would tend to have been represented in greater proportion in the extreme tail of the SG distribution.

There are no gold standards in terms of acceptable fit, but the adjusted $R^2$ associated with each model compared favourably with those achieved in the MVH main survey. The VAS median model achieved an adjusted $R^2$ of 0.96 compared to 0.97 in the MVH survey, and the VAS mean model achieved 0.68 compared to 0.47. The main MVH survey did not use SG, but in comparison with its TTO models the adjusted $R^2$ of the median models was 0.88 to 0.97 and 0.49 to 0.46 for the mean models.

The results confirm the wisdom of estimating coefficients for each dimension level, rather than assuming equal intervals between levels with dimension. The models selected for generating the tariffs (Tables 1, 2, 3 and 4a) were, however, all additive. It has been found in transport and marketing research using stated preference models that the main effects explain 80% or more of the variation in stated preference data (Louviere, 1988; Permain et al., 1991). Departures from an additive model are rare in this type of modelling work (Bates, 1988). Nonetheless, it was important to test for the existence of interactions. The limited testing possible with these data sets suggests there were no strong independent interaction effects. Where they were found to be significant, they displaced the main effect, and this was probably the result of collinearity. Similar results were found in the MVH main study, where the inclusion of interactions was also associated with anomalies in the models. The exception in the MVH work was an additional dummy variable for when the most severe level occurred within any dimension (i.e. 'N3'). The equivalent term in these models was not significant (Appendix).

### 8.3.2 Comparison with the MVH EQ-5D model

The Sheffield and MVH surveys used the same version of VAS, but direct comparisons are limited because the respondents to the surveys were not comparable. The MVH sample was a representative sample of the UK general population. The possible impact of this should be borne in mind in the comparisons below. Different estimation techniques were also used, since the Sheffield model was based on a fixed-effects model while the MVH models were estimated using random effects. However, this was shown in Section 8.3.3 to make little difference to the model estimates.

The median VAS model for the SF-6D estimated from the Sheffield data achieved a similar fit to the MVH model for the EQ-5D, with an adjusted R-squared of 0.97 compared to 0.96. The Sheffield individual level VAS model, however, was able to explain more of the variance with an adjusted R-squared of 0.68 compared to 0.47. This result suggests that the larger classification system of the SF-6D was able to explain more of the variance in VAS data.

A comparison of the coefficients on the dimension levels of the EQ-5D and SF-6D VAS models was more difficult since the dimensions, dimension levels and their 'vocabulary' are different. Some dimension levels appear comparable such as: 'extreme pain or discomfort' (EQ-5D) versus 'severe or very severe bodily pain' (SF-6D); 'moderate pain or discomfort' versus 'moderate bodily pain'; and 'some problems with washing or dressing self' versus 'limitation in bathing or dressing'. Comparisons of model coefficients are limited further by differences in the specification of the models. The MVH EQ models are also additive but include an extra term for when the most severe level occurs in at least one dimension (N3) that was not found to be significant in the SF-6D modelling work. In the MVH results the N3 term had a large and significant coefficient (i.e. 0.215). The marginal impact of the severe level of an EQ dimension is therefore dependent on the levels of the other dimensions.

A more direct comparison of these models is possible in terms of their ability to predict VAS in independent samples and this is presented in the next chapter.

### 8.3.3 Comparisons between VAS and SG models

There are substantial differences between the VAS and SG models[6]. The largest decrements in the VAS models were associated with the physical functioning dimension. The worst levels in this dimension, that is levels 5 and 6, were twice the size of the worst levels of pain and mental health. Whereas in the SG model, the worst levels of mental health and pain had the largest decrements and the physical functioning levels are of less relative importance. The most severe levels of the dimensions of role, social and vitality were moderately sized in the VAS models, but in the SG models are less important. None of the less severe levels of these dimensions was significant in the SG models.

Differences between SG and VAS valuations are well established in the literature (see Chapter 7). What has not been considered in previous studies are the differences in the relative value of health dimensions between valuation techniques. There has been no attempt to explain any differences that do emerge. The differences found here could reflect the argument that SG values more accurately reflect people's preferences. The SG question asks respondents to make a sacrifice and hence makes people think about the value of different dimensions of health. The responses to the SG question focus on pain and mental health, since these are more important in terms of their value in peoples lives. Avoiding severe pain or depression is worth far more to these respondents in terms of a sacrifice in expected survival than limitations in bathing or dressing. In contrast, it has been suggested that VAS is a measure of health in terms of such concepts as fitness rather than a reflection of its value (Chapter 7). The common perception of health is a physical one, and hence this dimension tends to dominate in the VAS model.

---

[6] To check this was not due to the patient respondents in the VAS data, these models were re-run on non-patient VAS data. The result was a model virtually identical in terms of fit and size of coefficients (Table A8.7)).

### 8.3.4 Implications for design of SF-6D

The estimated coefficients on the models indicate respondents were able to distinguish between most levels of each dimension of the SF-6D. The coefficients were usually in agreement with the ordinality of the scale. However, there were a number of inconsistencies. These may have been caused by multi-collinearity between the dimension levels in the model, misunderstandings of the SF-6D by respondents, or errors in the design of the SF-6D. A common inconsistency with the SF-6D was a positive coefficient on S2 (i.e. level two of social functioning). It seems the second level of social functioning was not regarded as negative in any of the models, and this suggests it should be excluded from the scale (i.e. merged with level one). Other common inconsistencies were between physical 4 and 5, pain 2 and 3, and vitality 3 and 4. It may have been that respondents were not able to distinguish between 'very mild' and 'mild' pain, nor between limitations in walking half a mile or 100 yards, yet these levels have an unambiguous ranking. Respondents were either unable to appreciate these distinctions in the context of the overall states, the valuation methods were too unreliable, or too small to impact on the model. A more understandable inconsistency arose was between 'a good bit of the time' and 'most of the time' on the vitality dimension (and to a less marked extent on the mental health dimension). These statements do not have an obvious ranking, and may indicate a need to merge these statements. Other inconsistencies were specific to the valuation method. The inconsistencies between mental 4 and 5 and pain 5 and 6 appeared in the VAS models but not the SG models. In contrast, there is little difference from zero in the coefficients on any levels of vitality in the SG. An energy dimension was also found to have no significant impact on EQ health state valuations (Gudex, 1991).

There was evidence of respondents having difficulty understanding parts of the SF-6D within the context of the health states, and this could be due to the size and complexity of SF-6D. Asking respondents to value whole health states, rather than single dimensions is extremely demanding of their cognitive abilities. Respondents ability to identify and value small differences in health states defined by this classification,

particularly using techniques such as SG, may be limited. In similar work, the MVH group reported very few inconsistencies using the much simpler EQ-5D. The question is whether the SF-6D was too large for respondents in terms of its number of dimensions and levels.

There could be a case for reducing the size of the SF-6D. Some levels such as S2, for example, could be excluded with no loss of information. The SG models suggest it might be appropriate to exclude all or most of the vitality dimension. This process of further refining the SF-6D could result in an instrument closer in size to the EQ-5D. However, the content of the dimensions will continue to differ in at least two respects. Firstly, it utilises the richer language of the SF-36. Secondly, it would be based more explicitly on respondents valuations and therefore the content will better reflect people's preferences than one generated by experts. There is already evidence from the SG model, for example, that physical functioning and pain require more levels than are available on the EQ-5D, whereas usual activities in terms of role and social may require fewer. The extra sensitivity will be retained where it is most important in terms of peoples preferences. There is no *a priori* reason why the three levels per dimension of the EQ-5D is the optimum balance.

## 8.4 Conclusion

In this chapter, methods were found for overcoming the problems of statistical estimation with pooled panel data and a number of alternative model specifications were explored. A set of four models was selected on the basis of goodness of fit, consistency with the SF-6D and parsimony for generating single index values for health from the SF-36. The resultant models, though additive, were able to explain most of the variation in VAS scores at the aggregate and individual levels and in SG values at the aggregate level. These three models also passed the standard diagnostic tests. The individual SG model explained nearly 50% of variation, but there were problems in terms of non-normality in the error terms and heterogeneity. Similar problems were encountered in the valuation of the modelling of health state values for the EQ-5D from TTO data.

It was reassuring to demonstrate the robustness of all four models with a comparatively small set of data. The task of valuing the SF-6D by statistical inference has been accomplished. This is only the second time that modelling work of this type has been undertaken with a health classification system[7], and the first study to model directly elicited SG data. The models provide algorithms for deriving a preference-based measure of health from the SF-36 and these are applied in the next chapter.

[7] This modelling work was undertaken concurrently but independently of the MVH Group at York.

**Table 8.1: Median VAS model**

| (1) Full version | | | (2) Consistent version | | |
|---|---|---|---|---|---|
| Variable | B | T | Variable | B | T |
| PH2 | -.039 | -1.4 | PH2 | -.074 | -3.0** |
| PH3 | -.099 | -3.5** | PH3 | -.120 | -4.6*** |
| PH4 | -.158 | -4.6*** | PH45 | -.176 | -6.0*** |
| PH5 | -.153 | -3.6*** | PH6 | -.254 | -6.6*** |
| PH6 | -.260 | -6.2*** | R2 | -.135 | -5.1*** |
| R2 | -.154 | -3.8*** | S3 | -.023 | -.9 |
| S2 | .006 | .1 | S4 | -.080 | 2.7** |
| S3 | -.006 | -.1 | S5 | -.139 | -3.8*** |
| S4 | -.042 | -.9 | PAIN23 | -.022 | -.8 |
| S5 | -.202 | -3.5*** | PAIN456 | -.083 | -2.8** |
| PAIN2 | -.066 | -1.9 | M2 | -.058 | -2.4* |
| PAIN3 | -.022 | -.7 | M345 | -.138 | -5.7*** |
| PAIN4 | -.102 | -3.3*** | V23 | -.026 | -1.0 |
| PAIN5 | -.102 | -2.1* | V4 | -.086 | -2.1* |
| PAIN6 | -.039 | -.7 | V5 | -.122 | -2.4* |
| M2 | -.083 | -3.2** | Constant | .932 | 31.9*** |
| M3 | -.192 | -5.3*** | | | |
| M4 | -.198 | -4.9*** | | | |
| M5 | -.047 | -.7 | | | |
| V2 | .001 | .0 | | | |
| V3 | .019 | .6 | | | |
| V4 | -.032 | -.7 | | | |
| V5 | -.121 | -2.3* | | | |
| Constant | .929 | 31.8*** | | | |
| df | 33 | | 41 | | |
| Adj $R^2$ | 0.960 | | 0.956 | | |
| F | 58.67 | | 82.53 | | |
| Normality | NS | | NS | | |
| Het. | NS | | NS | | |
| RESET test | NS | | NS | | |

**Table 8.2: Individual VAS model**

| a) Full version | | | b) Consistent version | | |
|---|---|---|---|---|---|
| | Coefficient | T | | Coefficient | T |
| PH2 | -.071 | -4.0*** | PH2 | -.085 | -5.8*** |
| PH3 | -.111 | -6.2*** | PH3 | -.120 | -7.2*** |
| PH4 | -.169 | -7.8*** | PH4 | -.172 | -8.8*** |
| PH5 | -.216 | -8.0*** | PH5 | -.232 | -10.5*** |
| PH6 | -.278 | -10.6*** | PH6 | -.279 | -12.4*** |
| R2 | -.106 | -4.2*** | R2 | -.088 | -5.1*** |
| S2 | .022 | .8 | S3 | -.033 | -2.0* |
| S3 | -.014 | -.5 | S45 | -.094 | -5.0*** |
| S4 | -.058 | -1.9** | PAIN23 | -.017 | -1.1 |
| S5 | -.113 | -3.1* | PAIN4 | -.082 | -3.0*** |
| PAIN2 | -.042 | -2.0 | PAIN56 | -.120 | -4.4*** |
| PAIN3 | -.011 | -.6 | M2 | -.070 | -4.6*** |
| PAIN4 | -0.89 | -4.6*** | M3 | -.115 | -5.7*** |
| PAIN5 | -.142 | -4.8*** | M45 | -.125 | -5.8*** |
| PAIN6 | -.117 | -3.3** | V2 | -.036 | -2.2* |
| M2 | -.079 | -4.8*** | V34 | -.074 | -3.5*** |
| M3 | -.134 | -5.8*** | V5 | -.100 | -3.4*** |
| M4 | -.148 | -5.8*** | Constant | .425 | 25.8*** |
| M5 | -.058 | -1.4 | | | |
| V2 | -.020 | -1.1 | | | |
| V3 | -.067 | -3.0** | | | |
| V4 | -.058 | -2.1* | | | |
| V5 | -.108 | -3.3*** | | | |
| Constant | .421 | 32.0*** | | | |
| | | | | | |
| df | 1333 | | 1339 | | |
| Adj. $R^2$ | 0.682 | | 0.682 | | |
| F | 127.6 | | 171.9 | | |
| Normality | NS | | NS | | |
| Het. | NS | | NS | | |
| RESET test | NS | | NS | | |

**Table 8.3: SG median model**

| (1) Full model | | | (2) Consistent version | | |
|---|---|---|---|---|---|
| | Coefficient | T | | Coefficient | T |
| PH2 | -.028 | -1.3 | PH2 | -.003 | -.1 |
| PH3 | -.038 | -1.7 | PH3 | -.029 | -1.4 |
| PH4 | -.094 | -3.8*** | PH456 | -.075 | -4.1*** |
| PH5 | -.104 | -3.3** | S3 | -.017 | -.9 |
| PH6 | -.090 | -3.1** | S45 | -.027 | -1.5 |
| R2 | .031 | 1.1 | PAIN2 | -.016 | -.7 |
| S2 | .012 | -.4 | PAIN34 | -.038 | -2.1* |
| S3 | -.037 | -1.2 | PAIN5 | -.136 | -4.4*** |
| S4 | -.093 | -2.7* | PAIN6 | -.195 | -5.0*** |
| S5 | .031 | .7 | M234 | -.025 | -1.4 |
| PAIN2 | .002 | .1 | M5 | -.218 | -6.0*** |
| PAIN3 | -.029 | -1.3 | V2 | -.005 | -.3 |
| PAIN4 | -.026 | -1.1 | V345 | -.020 | -8 |
| PAIN5 | -.129 | -3.9*** | Constant | -.1.025 | -52.3*** |
| PAIN6 | -.212 | -4.7*** | | | |
| M2 | -.026 | -1.3 | | | |
| M3 | .003 | .1 | | | |
| M4 | .036 | 1.2 | | | |
| M5 | -.301 | -6.3*** | | | |
| V2 | .026 | -1.2 | | | |
| V3 | .052 | -2.0 | | | |
| V4 | -.025 | -.8 | | | |
| V5 | -.011 | -.3 | | | |
| Constant | 1.028 | 46.3*** | | | |
| df | 33 | | 43 | | |
| Adj. $R^2$ | 0.8968 | | 0.876 | | |
| F | 22.15 | | 31.527 | | |
| Normality | NS | | NS | | |
| Het. | NS | | NS | | |
| RESET test | ** | | NS | | |

**Table 8.4a: Individual SG model**

| 1) Full version | | | 2) Consistent version | | |
|---|---|---|---|---|---|
| | Coefficient | T | | Coefficient | T |
| PH2 | -.014 | -.9 | PH2 | -.012 | -.9 |
| PH3 | -.027 | -1.7 | PH3 | -.028 | -2.1* |
| PH4 | -.070 | -4.0*** | PH45 | -.064 | -4.6*** |
| PH5 | -.063 | -2.8** | PH6 | -.098 | -5.4*** |
| PH6 | -.103 | -4.9*** | R2 | -.025 | -1.8 |
| R2 | -.037 | -1.7 | S3 | -.022 | -1.7 |
| S2 | .015 | .6 | S45 | -.033 | -2.5* |
| S3 | -.012 | -.5 | PAIN23 | -.022 | -1.6 |
| S4 | -.030 | -1.2 | PAIN4 | -.024 | -1.8 |
| S5 | -.028 | -.9 | PAIN5 | -.129 | -6.0*** |
| PAIN2 | -.028 | -1.4 | PAIN6 | -.163 | -5.7*** |
| PAIN3 | -.026 | -1.6 | M2 | -.021 | -1.9 |
| PAIN4 | -.027 | -1.6 | M34 | -.032 | -2.3* |
| PAIN5 | -.126 | -5.3*** | M5 | -.193 | -8.0*** |
| PAIN6 | -.164 | -5.1*** | V345 | -.020 | -1.8 |
| M2 | -.027 | -2.0 | Constant | -.150 | -11.3*** |
| M3 | -.043 | -2.3* | | | |
| M4 | -.031 | -1.5 | | | |
| M5 | -.200 | -5.8*** | | | |
| V2 | .012 | .8 | | | |
| V3 | -.013 | -.7 | | | |
| V4 | -0.001 | -.0 | | | |
| V5 | -.005 | -.20 | | | |
| Constant | .149 | 9.3*** | | | |
| df | 1013 | | 1021 | | |
| Adj.$R^2$ | 0.492 | | 0.495 | | |
| Normality | *** | | *** | | |
| Het. | *** | | *** | | |
| RESET test | NS | | NS | | |

**Table 8.4b: Individual logit SG model**

| 1) Full version | | | | 2) Consistent version | | |
|---|---|---|---|---|---|---|
| Attribute level | Coefficient | T | | | Coefficient | T |
| PH2 | -.400 | 1.9 | | PH2 | -.420 | -2.4* |
| PH3 | -1.135 | 5.3*** | | PH3 | -1.168 | -5.9*** |
| PH4 | -1.308 | 5.4*** | | PH4 | -1.281 | -5.7*** |
| PH5 | -1.452 | 4.8*** | | PH5 | -1.364 | -4.7*** |
| PH6 | -1.623 | 5.7*** | | PH6 | -1.486 | -5.7*** |
| R2 | -.649 | 2.2* | | R2 | -.843 | -4.3*** |
| S2 | -.299 | .9 | | S3 | -.174 | -.9 |
| S3 | -.404 | 1.3 | | S4 | -.305 | -1.4 |
| S4 | -.603 | 1.8 | | S5 | -.326 | -.8 |
| S5 | -.585 | 1.4 | | Pain 23 | -.071 | -.4 |
| Pain 2 | -.244 | .9 | | Pain 4 | -.301 | -1.5 |
| Pain 3 | -.141 | .6 | | Pain 5 | -1.017 | -3.3*** |
| Pain 4 | -.469 | 2.1* | | Pain 6 | -1.140 | -2.9** |
| Pain 5 | -1.153 | 3.6*** | | M2 | -.232 | -1.3 |
| Pain 6 | -1.253 | 2.9** | | M3 | -.321 | -1.5 |
| M2 | -.222 | 1.2 | | M4 | -.878 | -3.3** |
| M3 | -.303 | 1.2 | | M5 | -1.388 | -3.1** |
| M4 | -.905 | 3.2** | | V2 | -.240 | -1.2 |
| M5 | -1.453 | 3.1** | | V345 | -.465 | -1.9 |
| V2 | -.185 | .9 | | constant | 2.641 | -13.3*** |
| V3 | -.473 | 1.9 | | | | |
| V4 | -.194 | .6 | | | | |
| V5 | -.182 | .5 | | | | |
| constant | -2.740 | 12.7*** | | | | |
| | | | | | | |
| df | 1013 | | | 1017 | | |
| Adj. R2 | 0.422 | | | 0.423 | | |
| F | 33.9 | | | 40.9 | | |
| Normality test | *** | | | *** | | |
| Het. test | *** | | | *** | | |
| RESET test | *** | | | *** | | |

**Table 8.5a: Comparison of estimated with actual median values - VAS data**

| Health states | n | Median Estimate | SE Estimate | Actual Median | Estimated - Actual median |
|---|---|---|---|---|---|
| 111212 | 35 | .89 | .02 | .89 | .00 |
| 111311 | 55 | .93 | .03 | .90 | .03 |
| 111312 | 31 | .89 | .02 | .90 | -.01 |
| 124143 | 189 | .60 | .02 | .55 | .05 |
| 211111 | 34 | .88 | .03 | .90 | -.03 |
| 222432 | 32 | .47 | .03 | .45 | .02 |
| 224244 | 66 | .43 | .03 | .40 | .03 |
| 311211 | 30 | .80 | .03 | .72 | .08 |
| 311222 | 34 | .69 | .03 | .70 | .00 |
| 313333 | 31 | .60 | .03 | .65 | -.05 |
| 322323 | 33 | .55 | .03 | .65 | -.10 |
| 323422 | 30 | .49 | .03 | .44 | .05 |
| 422413 | 35 | .51 | .02 | .58 | -.06 |
| 422434 | 22 | .31 | .03 | .21 | .10 |
| 423122 | 31 | .50 | .03 | .53 | -.03 |
| 521412 | 21 | .51 | .02 | .53 | -.01 |
| 523111 | 22 | .61 | .03 | .63 | -.01 |
| 525112 | 55 | .48 | .03 | .42 | .06 |
| 525555 | 21 | .15 | .03 | .12 | .03 |
| 623424 | 133 | .25 | .02 | .30 | -.05 |
| 624415 | 30 | .07 | .03 | .21 | .05 |
| 624645 | 21 | .12 | .03 | .11 | .01 |
| 625555 | 31 | .04 | .02 | .06 | -.02 |

**Table 8.5b: Comparison of estimated with actual mean - VAS data**

| Health states | n | Mean Estimate | SE Estimate | Actual Mean | Estimated - Actual Mean |
|---|---|---|---|---|---|
| 111212 | 35 | .87 | .01 | .89 | -.02 |
| 111311 | 55 | .91 | .02 | .88 | .03 |
| 111312 | 31 | .87 | .01 | .90 | -.03 |
| 124143 | 189 | .54 | .01 | .54 | .00 |
| 211111 | 34 | .84 | .02 | .87 | -.03 |
| 222432 | 32 | .52 | .02 | .48 | .04 |
| 224244 | 66 | .44 | .02 | .42 | .02 |
| 311211 | 30 | .79 | .02 | .71 | .08 |
| 311222 | 34 | .68 | .02 | .67 | .01 |
| 313333 | 31 | .56 | .02 | .61 | -.05 |
| 322323 | 33 | .55 | .02 | .62 | -.06 |
| 323422 | 30 | .49 | .02 | .46 | .03 |
| 422413 | 35 | .51 | .02 | .55 | -.04 |
| 422434 | 22 | .39 | .02 | .27 | .12 |
| 423122 | 31 | .52 | .02 | .55 | -.03 |
| 521412 | 21 | .48 | .02 | .50 | .02 |
| 523111 | 22 | .57 | .02 | .63 | -.06 |
| 525112 | 55 | .47 | .02 | .44 | .03 |
| 525555 | 21 | .16 | .02 | .16 | .00 |
| 623424 | 133 | .30 | .01 | .36 | -.06 |
| 624415 | 30 | .28 | .02 | .22 | .06 |
| 624645 | 21 | .12 | .01 | .11 | .01 |
| 625555 | 31 | .12 | .01 | .07 | .05 |

**Table 8.6a: Comparison of estimated with actual median values - SG data**

| Health State | n | Estimated Median | SE Estimate | Actual Median | Estimated - Actual Median |
|---|---|---|---|---|---|
| 111212 | 20 | 1.00 | .02 | .99 | .01 |
| 111311 | 44 | .99 | .02 | .99 | .00 |
| 111312 | 18 | .98 | .02 | .98 | .00 |
| 124143 | 127 | .95 | .01 | .95 | .00 |
| 211111 | 21 | 1.02 | .03 | .99 | .03 |
| 222432 | 31 | .95 | .02 | .96 | -.01 |
| 224244 | 50 | .94 | .02 | .95 | -.01 |
| 311211 | 29 | .98 | .02 | .98 | .00 |
| 311222 | 22 | .95 | .02 | .96 | -.01 |
| 313333 | 18 | .90 | .02 | .85 | .05 |
| 322323 | 22 | .91 | .02 | .90 | .01 |
| 323422 | 30 | .91 | .02 | .95 | -.04 |
| 422413 | 22 | .89 | .02 | .85 | .04 |
| 422434 | 22 | .87 | .02 | .94 | -.07 |
| 423122 | 18 | .90 | .02 | .85 | .05 |
| 521412 | 18 | .91 | .02 | .90 | .01 |
| 523111 | 22 | .93 | .02 | .95 | -.02 |
| 525112 | 40 | .92 | .02 | .96 | -.04 |
| 525555 | 22 | .55 | .02 | .50 | .05 |
| 623424 | 82 | .85 | .01 | .85 | .00 |
| 624415 | 30 | .86 | .02 | .85 | .01 |
| 624645 | 20 | .68 | .04 | .70 | -.02 |
| 625555 | 28 | .55 | .02 | .60 | -.05 |

**Table 8.6b: Comparison of estimated with actual mean values - SG data**

| Health State | Mean Estimate | SE Estimate | Actual Mean | Estimated - Actual Mean |
|---|---|---|---|---|
| 111212 | .97 | .01 | .96 | .01 |
| 111311 | .97 | .01 | .99 | -.02 |
| 111312 | .97 | .01 | .96 | .01 |
| 124143 | .88 | .01 | .88 | .00 |
| 211111 | .98 | .02 | .97 | .01 |
| 222432 | .90 | .02 | .92 | -.02 |
| 224244 | .85 | .02 | .88 | -.03 |
| 311211 | .94 | .01 | .97 | -.03 |
| 311222 | .92 | .01 | .92 | .00 |
| 313333 | .87 | .02 | .82 | .05 |
| 322323 | .88 | .02 | .91 | -.03 |
| 323422 | .87 | .01 | .91 | -.04 |
| 422413 | .86 | .01 | .83 | .03 |
| 422434 | .83 | .01 | .89 | -.06 |
| 423122 | .86 | .01 | .80 | .06 |
| 521412 | .88 | .01 | .83 | .05 |
| 523111 | .88 | .01 | .89 | -.01 |
| 525112 | .87 | .01 | .90 | -.03 |
| 525555 | .53 | .02 | .51 | .02 |
| 623424 | .78 | .01 | .78 | .00 |
| 624415 | .79 | .01 | 83 | -.04 |
| 624645 | .62 | .02 | .62 | .00 |
| 625555 | .50 | .02 | .54 | -.04 |

**Table 8.7: Correlation matrix of the SF-6D dimensions**

| | Role | Social | Pain | Mental | Vitality |
|---|---|---|---|---|---|
| **Physical** | .42 | .37 | .59 | -.07 | .47 |
| **Role** | | .72 | .23 | .42 | .56 |
| **Social** | | | .06 | .60 | .61 |
| **Pain** | | | | .12 | .53 |
| **Mental** | | | | | .58 |

**Figure 8.1: Multi-level data structure**

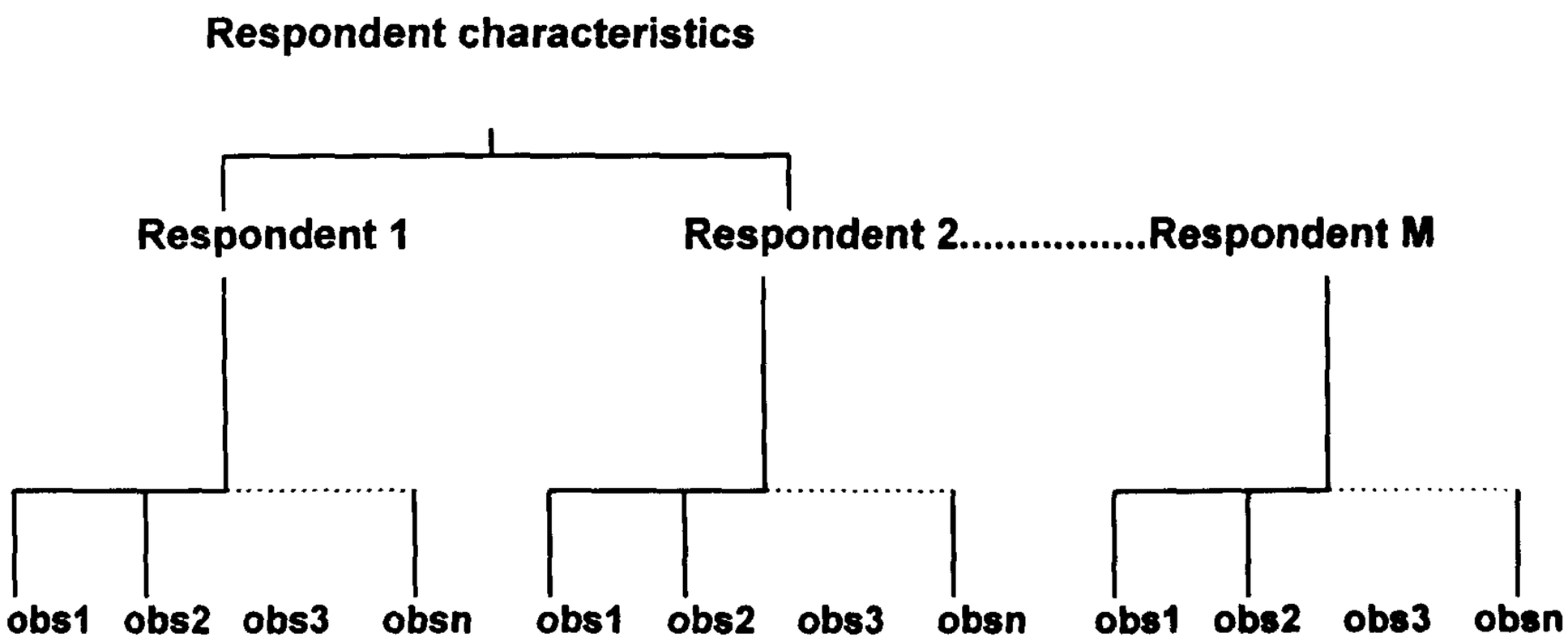
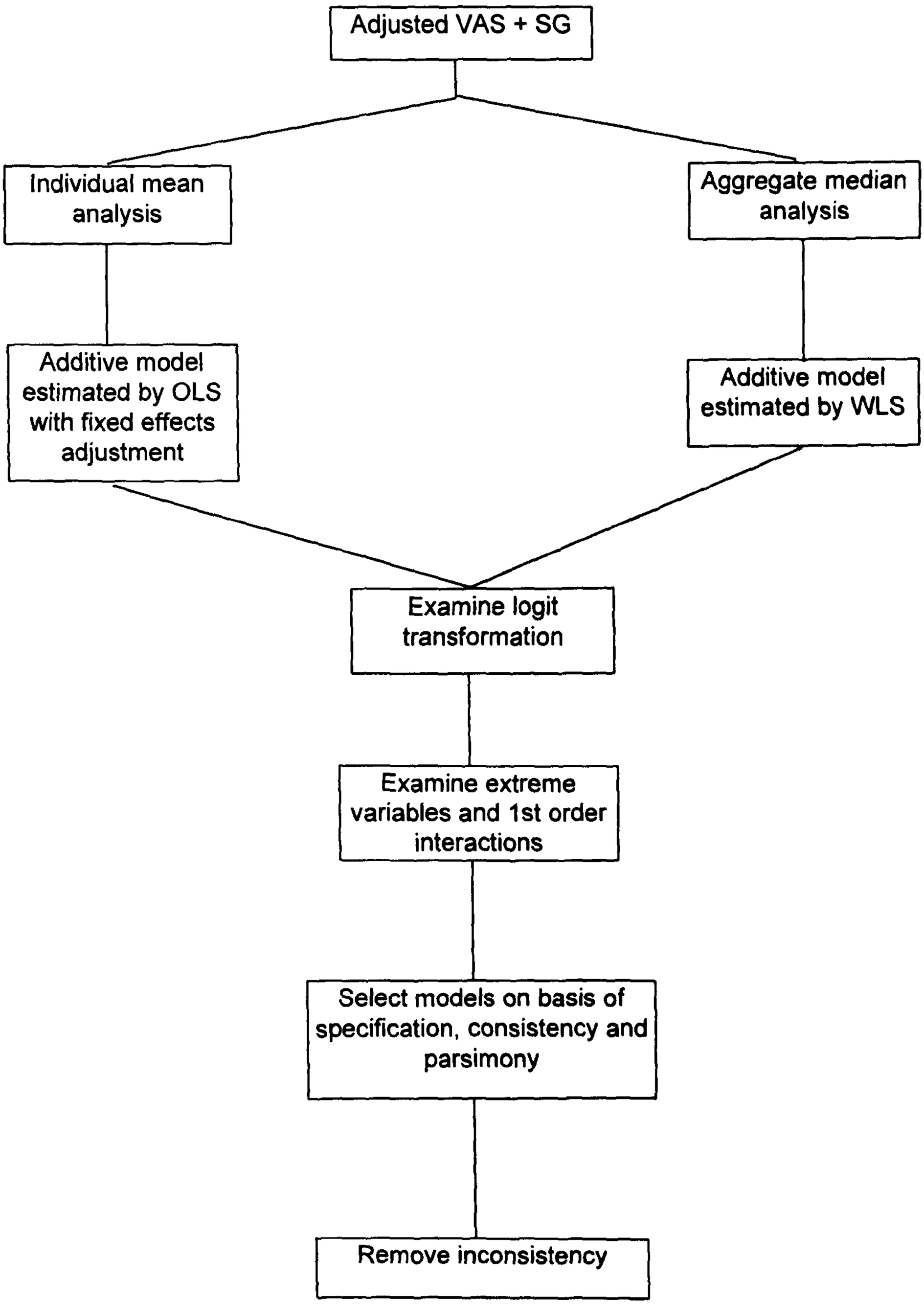


**Figure 8.2: Analysis plan**



# Chapter 9

# Applications

The models presented in the previous chapter provide the means for generating preference-based measures of health from SF-36 data. In this chapter, these models are applied to five data sets. The primary purpose is to examine the extent to which the adaptation of the SF-36 into the SF-6D and the further simplifications brought about by the modelling (i.e. the merging of dimension levels) has reduced the sensitivity of the original instrument to health differences and changes.

The chapter begins by a description of methods including the analyses, the choice of algorithms and the SF-36 data sets. This is followed by a presentation and discussion of the results for the reliability, descriptive validity, and empirical validity of the values generated from the SF-6D, and a comparison with the EQ-5D. The final application is a cost-utility analysis using the results of a randomised clinical trial of alternative treatments for inguinal hernia patients. The final section considers the implications of these findings.

## 9.1 Methods

### 9.1.1 Analyses

This chapter presents analyses of the measurement properties of the preference-based health state values derived from the SF-36 data using the models estimated in the last chapter. These analyses are based on the criteria described in Chapter 3 of reliability, descriptive validity and empirical validity. The performance of the SF-6D has been compared to the EQ-5D against these criteria. Finally, the new preference-based measure is used to undertake a CUA.

Descriptive validity

The descriptive validity is assessed in terms of construct validity and repeats analyses undertaken on the SF-36 dimensions on five data sets: a general population adult population (18-74), a female elderly population (over 74), patients attending clinics with

diagnosed chronic obstructive pulmonary disease and osteoarthritis of the knee (Brazier et al., 1992; Brazier et al., 1996a; Brazier et al.,1996b; Harper et al., 1997) and patients recruited into a trial for hernia repair. These analyses will examine the extent to which the adaptation of the SF-36 into the SF-6D and the further simplifications brought about by the modelling has reduced its sensitivity.

The specific analyses are as follows:

1) Construct validity will be assessed using the general population data sets in terms of the pattern of health state values by age, socio-demographic group and use of services. These constructs include an inverse relationship between health and age, with the exception of mental health where a non-linear relationship has been suggested (Kessle et al., 1992); professional and managerial groups should report better health than those in manual employment (GHS, 1988) and this pattern would be expected to be repeated with the SF-6D health measures; and people who have recently used health services, including hospital stay, attendance at hospital outpatient clinic, or consulting with general practitioners, would be expected to have poorer health than non-users across a general population.

2) For the patient data sets, construct validity is examined against the severity of the condition, comorbidity and recent hospital admission.

<u>Empirical validity</u>

Empirical validity has been examined in terms of the degree of convergence between the VAS values derived from the EQ-5D and the SF-6D and the patient's own stated preferences elicited by VAS. This is possible because the EQ instrument includes a VAS question completed by the patients themselves. This comparison has been undertaken in terms of correlation and mean differences.

Empirical validity has also been examined against hypothetical preferences. In two of the data sets, patients were asked at each follow-up whether their health in general had changed compared with the last time they completed the questionnaires. Specifically they were asked, "Would you rate your health in general now: much better, somewhat

better, about the same, somewhat worse, and much worse?'' Their answer to this question should be related to a change in the preference-based health values.

Comparison of SF-6D and EQ

These measures will be compared in terms of the health state values they estimate for the five data sets (i.e. general population, elderly female population, COPD patients, osteoarthritis patients, and patients awaitng hernia repair). EQ-5D will be valued using the published MVH algorithms from the VAS and TTO data. The health state values for these data sets will be correlated and mean differences calculated. At the very least, the correlation between the VAS scores derived from SF-6D and EQ-5D should exceed their correlation with the SG and TTO values.

The measures will also be compared in terms of their descriptive and empirical validity using the methods of analyses outlined above. Descriptive validity will also be compared by the distribution of responses across the dimensions of the EQ and SF-6D classification to examine 'floor' and 'ceiling' effects in the general population sample. These are important attributes of a scale, since patients at the floor of a scale, that is near the lowest possible point on a scale, cannot show a deterioration on that dimension and patients near the ceiling cannot improve. For a large number of patients at the floor and/or the ceiling the dimensions would indicate a limitation in the ability of the measure to describe change. The size of these 'floor' and 'ceiling' effects will be compared between SF-6D and Euroqol. This analysis repeats an earlier comparison of the SF-36 and the EQ-6D (Brazier et al., 1993).

A floor effect is important for the evaluation of inpatients and others with serious illness (Bindman et al., 1990), but it is less obvious whether a ceiling effect is important since these patients may not have health problems of significance. In the earlier comparison of SF-36 and EQ-6D, it was found that those who recorded no problem according to the EQ classification contained groups who could be defined as in 'better' or 'worse' health by the SF-36. This was confirmed in terms of the construct variables of age, sex, socio-economic class and use of services. Similar analysis will be undertaken for the SF-6D to examine whether the apparent advantages of the SF-36 over the EQ-6D are maintained.

Case study: a cost-utility analysis of alternative treatments for inguinal hernia

A cost-utility analysis (CUA) has been undertaken using the results of a clinical trial of laparoscopic versus open repair for inguinal hernia. The overall QALY gain associated with the hernia procedures has been estimated using the SG valuation of the SF-6D health states at each assessment, since SG values are believed to provide a better reflection of the quantity/quality trade-off required for estimating QALYs than VAS (Chapters 3 and 7). The number of QALYs generated is estimated as the area between the line joining up the SG health state values at each follow-up assessment and the baseline (Williams, 1985) using a formula recommended by Matthews and colleagues (1990) in the British Medical Journal[1]. This requires, *inter alia*, an assumption that the differences between the pre-operative and six month post-operative assessments are maintained for some time period and five years is assumed for this exercise.

### 9.1.2 The algorithms

The four models selected in the previous chapter for valuing the SF-6D provide algorithms for estimating preference-based values for the 9000 health states defined by the SF-6D. To obtain median health state values the application of estimated models is straightforward. The model predicts the value of each health state as the constant term minus the coefficients associated with each dimension level. The algorithm for calculating mean health state values is the same, except the individual level models in Chapter 8 predict the residuals of the respondent model, with a mean of zero. To predict health state values it is necessary to add a mean respondent effect to the constant term.

The median VAS value of health state 224244 can be calculated as follows (from Table 8.1):

---

[1] In the case of a baseline assessment $h_1$, and two follow-up assessments $h_2$ and $h_3$, at time intervals of $t_{12}$ and $t_{23}$ respectively, the formula for calculating the overall gain (or loss) per day is as follows: $[\frac{1}{2} t_{12} (h_1 + h_2) + \frac{1}{2} t_{23} (h_2 + h_3) - t_{13} * h1]/t_{13}$.
This has been shown to be a close approximation for the area between the line joining each assessment and the baseline (Matthews et al, 1990).

| | |
|---|---|
| Constant term | 0.932 |
| Physical 2 | - 0.074 |
| Role 2 | - 0.135 |
| Social 4 | - 0.080 |
| Pain 2 | - 0.022 |
| Mental 4 | - 0.138 |
| Vitality 4 | - 0.086 |
| Value of health state 224244 | 0.397 |

The constant term is the model estimate for health state 111111, and therefore should be equal to unity for both the VAS and SG models since the VAS scores were adjusted to ensure health state 111111 equals 1.0, and the SG values were obtained in gambles where this was one of the reference states and set to 1.0. The estimated constant terms and their 95% confidence intervals were, however, as follows: median VAS was 0.932 (0.87 to 0.99); mean VAS was 0.922 (0.89 to 0.96); Median SG was 1.025 (0.989 to 1.064) ; and mean SG was 0.994 (0.986 to 1.020). The differences from unity in the SG models could be due to chance. For VAS , the difference is statistically significant, and suggests the models are poor at predicting the value of health state 111111. A similar result was found in the main MVH survey, where the constant term was 0.845 for mean VAS model and 0.919 for the mean TTO model. The MVH group interpreted the constant term to imply that 'any move away from full health is associated with a substantial loss of utility' (MVH, 1995). In their algorithm for valuing the EQ-5D, full health is assigned a value of 1.0, and any ill health state automatically has its score reduced by the constant term (i.e.0.155).

One explanation for the difference from unity could be a discontinuity in the scale between ill health and full health. Given the reservations in the literature about the intended meaning behind VAS responses, the scale may not be continuous and is unlikely to have interval properties throughout its length, particularly near the two end points. This would suggest the MVH interpretation could be correct for the VAS scale (though this is not relevant for the TTO models). It also lends further support to the argument that VAS does not reflect people's quantity/quality trade-off and hence should not be used to estimate QALYs. However, this is not the correct statistical interpretation

of the constant. The algorithms derived from the SF-6D models used in this chapter adopt the usual interpretation of the constant as the intercept term and hence is the value of health state 111111.

There is an argument for adjusting the models estimated in the previous chapter according to the background characteristics of the patients on whom it is being applied. Some studies have found variations in health state values to be related to the socio-demographics and health status of the respondents (Froberg and Kane, 1989b), but this is not a consistent finding (see, for example, the valuation of the QWB in Kaplan and Bush, 1982). The MVH study was the largest survey of its kind and found age and gender to be significant variables in the TTO models, and educational attainment in the VAS models. The MVH team have estimated separate tariffs for the EQ-5D for groups defined by these variables.

To examine the extent to which background characteristics were important in the Sheffield valuation survey, the mean VAS and SG models were re-estimated (without the fixed-effects adjustment) with the addition of variables for gender, age group (18-44; 45-64; 65 and over), limiting long standing illness, general health rating, and whether the respondent was recruited from an outpatient clinic (only applicable to the VAS model). The only significant variable was age group, and this was limited to the VAS model (Table 9.1). Respondents over 65 were found on average to rate health states by an extra 0.073 (on a 0 to 1.0 scale) above the younger age groups. There could be a case for having a separate algorithm for this older group.

The MVH group leaves the choice to the user, and provides tariffs for different groups of the population. The SF-6D models, however, are based on comparatively small and unrepresentative data, and hence there can be little confidence in the findings for the background characteristics of respondents, particularly given that the result contradicted the MVH findings, where age was not significant in the VAS model. There would seem little justification, therefore, in adding to the complexity of the analyses presented in this chapter with extra models. Furthermore, for informing resource allocation decisions

there is an argument for using the values of the general population and this is the recommendation of the MVH group (Dolan et al., 1995).

As seen in Chapter 8, there was little difference in the mean and median health state values (see Table 9.4). Therefore in the cause of parsimony, only the mean health state values have been used in the results presented below.

### 9.1.3 Data sets

The algorithms for valuing the SF-6D have been applied to five SF-36 data sets. These data sets were chosen because they were accessible to the author in the raw form necessary to apply the algorithms.

1. General population

The UK SF-36 questionnaire and the EQ-6D were included in a postal survey of 1980 people aged 16-74 years randomly selected from two general practice lists in Sheffield (Brazier et al., 1992). The sociodemographic characteristics and use of health services of the 1582 respondents did not differ from those found in the General Household Survey for the same age range, except for socio-economic class where the sample included fewer people employed as managers, but more with intermediate and junior non-manual employment and more females in employment.

2. Chronic obstructive pulmonary disease

One hundred and fifty two adult patients clinically diagnosed with chronic obstructive pulmonary disease (COPD) attending routine appointments at a chest clinic were assessed at recruitment ( a response rate of 94%), and at six and twelve month follow-ups. At each assessment the SF-36 and EQ-5D were administered alongside two condition-specific measures of patient-perceived health, questions about breathlessness, and tests of exercise tolerance and respiratory function (Harper et al.,1997).

3. Elderly women

This is a sample of 380 women over 75 years of age who participated in a pilot study for a randomised clinical trial of the use of clodronate for the prevention of hip fractures (a

response rate of 97%). The SF-36 and EQ were administered, along with the OPCS disability survey instrument, at baseline and then again on a randomly selected sub-sample of respondents six months later (Brazier et al., 1996a).

4. Osteoarthritis of the knee

This sample contains two groups of patients with osteoarthritis of the knee: 112 patients recruited from rheumatology clinics (response rate 90%) and 118 recruited prior to a knee replacement procedure (response rate 79%). These patients received the SF-36, EQ-5D and two disease condition-specific questionnaires at recruitment and six months later (Brazier et al., 1996b).

5. Inguinal hernia

These were patients who had a primary, unilateral inguinal hernia and met the criteria for day surgery (Lawrence et al., 1995)[2]. In all, 130 patients were allocated randomly between open and laparoscopic surgery. General anaesthesia was administered to all patients. The SF-36 and EQ were administered 10 days pre-operatively and at 10 days, six weeks, three months and six months post-operatively. These health measures were collected alongside cost data in order to examine 'cost-effectiveness'. This study provides the only opportunity available at the time of writing the thesis to undertake a cost-utility analysis using preference-based values derived from the SF-36.

## 9.2 Results

### 9.2.1 Reliability

Re-test reliability results for the COPD patients are presented for the SF-6D and EQ-5D values over two time periods: between initial assessment to six months and six months to a year. For both periods, the rank correlation coefficients were significant between test and re-test for each measure (Table 9.2). The coefficients were similar for the SF-6D and EQ-5D health state values, but rather lower for the SF-6D SG value in period

[2] Dr Lawrence and colleagues, based in Oxford, have kindly provided this data set.

one. The mean differences between test and re-test were not significantly different from zero for any of values. The 95% confidence intervals around the mean differences were within plus or minus 0.05, the exception being EQ-5D TTO values which had 95% confidence intervals of -.075 to .069 and -.139 to 0.46.

In those elderly female patients who said their health had not changed over a six month period, the test and re-test values of the SF-6D and EQ-5D were significantly correlated ($P<0.001$). The mean differences were non-significant and within a 95% confidence interval of plus and minus 0.05, except for the EQ-5D TTO (Table 9.3).

These results confirm the re-test reliability of the SF-6D, and its similar performance compared to the EQ-5D.

9.2.2 Descriptive validity

<u>The SF-6D health state values</u>

The distribution of VAS and SG values by age, socio-economic class, and use of health services conformed to the hypotheses described in the methods section (Table 9.4). The values significantly decreased by age and socio-economic class ($P<0.001$). Patients who consulted their GP in the previous two weeks, attended an outpatient clinic in the previous three months or were admitted as an inpatient in the last year had significantly lower VAS and SG values ($P<0.001$). These results repeat the findings for the SF-36 dimension scores (Brazier et al., 1992).

Similar results were also found in the elderly female population. SF-6D values were significantly different across six indicators of health: GP visit in last two weeks, outpatient and A & E attendances in last three months, inpatient stay in previous year, any long standing illness, and OPCS disability category (Table 9.5). The differences were significant for all four values at the 1% level, except for outpatient attendance in the last three months where the difference was only significant at the 5% level.

In the COPD patient sample, SF-6D values were found to be significantly related to three widely recognised indicators of the severity for this condition (Jones, 1991) and two other health variables (Table 9.6). The values were able to differentiate between patient groups defined in terms of severity of breathlessness, distance walked in the exercise tolerance test, hospital admission in the last year and comorbidity ($P<0.05$). A VAS rating of breathing difficulties taken at the end of the six minute walking test is known to be a weak indicator of health, nonetheless the differences in SF-6D values were in the right direction but failed to reach significance in part owing to small numbers. The absence of a significant difference by respiration function ($FEV_1$% predicted) is supported in the findings of previous studies (Jones, 1991). These results confirm results for the SF-36 dimension scores (Harper et al., 1997).

These results suggest the SF-6D retains the construct validity of the SF-36 in the general population samples and in the COPD patient group.

Comparison of SF-6D with EQ

An important question is whether the SF-6D retains the extra sensitvity of the SF-36 over the EQ. The sensitivity of these measures has been compared in terms of the distribution of responses of the adult general population across their dimension levels. The distribution of responses of comparable dimensions has been cross tabulated between the instruments (Table 9.7). The frequency distribution of EQ responses were found to be considerably more skewed than the SF-6D for comparable dimensions. The skewness reflects in part the limitation of having only two or three levels for each dimension of the EQ compared to five or more for five out of the six dimensions of the SF-6D. The percentage of responses at the 'ceiling' of the functional dimensions was over 95% for the EQ compared to 37-58% for SF-6D. The differences were less marked for emotional well-being and pain, with 81% on the ceiling of anxiety and depression (EQ) and 58% on mental health (SF-6D), and 64% on the pain dimension of EQ compared to 38% on the SF-6D. The EQ response of 'no problem' is associated with a large spread of SF-6D dimension categories.

The construct validity of this apparent extra sensitivity of the SF-6D has been confirmed in those respondents who were at the ceiling of the EQ dimensions. For those identified to have no problem according to the EQ-6D, the distribution of responses across the scales of the dimensions of the SF-6D were found to be significantly associated with age, GP visits, outpatient attendance and inpatient admissions in the majority of cases (Table 9.8). The exceptions were the role dimension, probably because it has only two levels in the SF-6D, and the lack of association between age and the pain and social dimensions. The other apparent anomaly was the percentage of inpatients, which was significant only for the physical and social dimensions.

The models used to value health states did not use all the dimension levels of the SF-6D. The analysis of skewness and discriminatory power has therefore been repeated for the reduced consistent versions of the VAS and SG models used to value SF-6D. The reduced dimension scales of the SF-6D have been cross-tabulated with comparable EQ-6D dimensions on Tables 9.9 and 9.10. The distribution of SF-6D dimensions continues to be less skewed than the EQ-6D within dimension, and the distribution of construct variables confirms the prior hypotheses in the majority of cases (Tables 9.11 and 9.12). Only two of the relationships between the construct variables and the dimension scales lost significance: those between the social dimension and the percentage using outpatient and inpatient services.

This evidence would suggest the SF-6D and the reduced version used in the final models have more scope for measuring health improvement than the EQ-6D. The importance of this sensitivity has been confirmed by the construct validation. However, the EQ-6D has since been replaced by the EQ-5D, and the number of levels has been increased to three for all dimensions. This may have reduced the insensitivity of the EQ but it is unlikely to have removed it entirely.

The sensitivity of SF-6D against EQ-5D has been compared in the COPD patient sample. The EQ-5D health (TTO and VAS) values were significantly different between patient groups defined by breathlessness and exercise tolerance (Table 9.6). However, patients with comorbidities did not have lower scores in contrast to the SF-6D (VAS

and SG) values. In the elderly female data set, the differences in EQ-5D values were significant across hospital stay, long standing illness and OPCS category (Table 9.5). The mean differences by GP visits were not as significant statistically as for the SF-6D, and not significant at all for A & E attendance.

These apparent differences in the sensitivity of the SF-6D over the EQ-5D must be interpreted with care, since the differences could be explained by the time frames of the instruments. SF-36 questions ask about health over the last four weeks (or one week in the acute version), whereas the EQ-5D asks about today. As a result, the SF-36 is likely to be more sensitive in terms of recent health service use, such as visits to the GP in the last two weeks and to a lesser extent a hospital attendance in the last three months, since this reflects past rather than current health. Further work is required in order to ascertain the relative importance of the time frame effect. It could explain some of the apparent differences in sensitivity in terms of descriptive validity between the measures.

### 9.2.3 Empirical validity

Stated preferences: convergence with own VAS ratings

The SF-6D and EQ-5D were administered to patients alongside the EQ instruments 'own health' VAS in the COPD, osteoarthritis of the knee, elderly female and the hernia repair studies. The VAS estimates derived from the SF-6D and EQ-5D were found to be significantly correlated to the patient's own VAS rating (Table 9.13). The SF-6D VAS estimates had on average a higher correlation with self-rating than EQ-5D VAS, but the difference is small (i.e. 0.53 vs. 0.47) and given the variation across conditions it could be due to chance.

Both of the estimated VAS values were significantly less than patients' own rating (Table 9.13). This difference was considerably larger for the SF-6D than the EQ-5D. The plots of these mean differences against derived values indicate that for each patient group, the mean difference is positively correlated with the derived values (Appendix 7, Plots A7.1a-7.4b). This was confirmed by the significant correlations of the difference

between the estimated VAS values and patients' own VAS rating in both instruments for all patient groups except OA (Table 9.14).

It was disappointing to find that the estimated VAS values were only moderately correlated to patients own rating, and that they were consistently lower than the patients' own rating. Furthermore, this differential was larger for the SF-6D. There are a number of explanations for these findings. Underestimating the patients' own VAS ratings could partly be explained by the different characteristics of the respondents. The COPD, elderly female and osteoarthritis patients were significantly older than respondents in the MVH or Sheffield survey. The patient populations had 57%, 100% and 73% respectively over 65 compared with 24% and 13%. Age was not found to be significant in the MVH model, but accounted for an 0.073 increase in SF-6D VAS health state valuation by the over 65s. The more elderly patients are therefore likely to value the same states more highly. This explanation is also consistent with the fact that the ratings of younger hernia patients are nearer to the estimated VAS values. Furthermore, three of the patient populations also had higher proportions in the manual occupations. These occupational groups were found in the MVH study to rate EQ-5D health states more highly.

A more general explanation comes from the evidence in the literature that people experiencing a health state tend to value it more highly than members of the general population imagining the health state (Sackett and Torrance, 1978). It has been suggested that people adapt to a health state and therefore can reduce its impact on their quality of life and adjust expectations to their circumstances.

The discrepancies are complicated further by the positive correlation of the difference (between the estimated VAS values and the patients' own ratings) and the estimated VAS value. An explanation of the general pattern could be the tendency for respondents to spread their ratings along the length of the scale (Stevens and Galenter, 1957). In the MVH and Sheffield surveys, respondents valued sets of health states and so tended to use the entire scale. Patients valuing their own health were valuing only one state and were likely to be drawn to the middle and indeed there was a

preponderance of self-ratings in the mid-range for all conditions. At low values, the patients' own VAS rating therefore exceeds the estimated value, and vice versa for high values. Given all four samples were patient groups with health problems, most were at the lower end of the health spectrum and therefore it is not surprising that the patients' own valuations exceed the estimated values.

This comparison has confirmed the ordinal validity of the SF-6D and EQ-5D values, but it cannot be regarded as a valid test of the empirical validity of the cardinal properties of these measures. The differences in the background characteristics of the respondents and the spreading effect make such a comparison of means inappropriate. Furthermore, there are more fundamental reasons for doubting whether VAS can be a cardinal measure of preferences. A comparison between estimated TTO and SG values from the EQ-5D and the SF-6D and patients' own TTO and SG valuations would have been a better test of empirical validity against stated preferences.

Hypothetical preferences

This has been examined against patient-perceived changes in health between assessment on a simple three point scale: better, same or worse. In the COPD patient group, the mean differences in SF-6D values between assessments were in the right direction in relation to this transition question, but only significant for the SF-6D VAS value (Table 9.15). Differences in three of the SF-36 dimension scores had been significant. The EQ-5D VAS and TTO values were not significantly related to health transition, and the difference was of the wrong sign in those patients who had reported that their health had improved. (N.B. differences in the patients' own VAS rating were significant and in the right direction.)

The reduction in responsiveness compared to some dimensions of the SF-36 may have been due to the adaptation of the SF-36 into the SF-6D and the further simplifications brought about by the modelling. Another explanation is that the derivation of a single index uses items from all the dimensions (excluding GHP) has the effect of pooling the responsive and unresponsive dimensions. In populations experiencing a change across all dimensions this attenuation may not occur, but where there is a differential effect

there will be an apparent reduction in responsiveness to change. Nonetheless the SF-6D VAS values were found to be more responsive than EQ-5D in the COPD study. The EQ classification seems less able to measure improvement in this patient group.

### 9.2.4 Comparison with EQ-5D

The values generated by the SF-6D are significantly correlated to the EQ-5D values across the five samples (Table 9.16). The largest correlations are between the values derived from the same classification, rather than the same valuation technique. The average correlation across the five conditions between EQ-5D VAS and EQ-5D TTO was 0.99 and between SF-6D VAS and SG was 0.86 compared to the average correlation between SF-6D VAS and EQ-5D VAS of 0.62. The correlations between the VAS estimates were no larger than those between SF-6D VAS and EQ-5D TTO or EQ-5D VAS and SF-6D SG. There were also important differences in the mean values of the different estimates (Table 9.17): SG values consistently exceeded TTO and VAS values and the SF-6D VAS index values were consistently less than EQ-5D VAS values.

The observed differences between SG and VAS health state values are the same as those found in comparisons of values elicited directly and the reasons for the differences have been discussed at length in Chapter 7. The result that SG values exceed TTO values also confirms findings from comparisons of values elicited by these techniques (Read et al., 1984; Wolfson et al.,1982). The findings provide some support for the empirical validity of using SF-6D and EQ-5D to derive estimates of these values.

The systematic differences between the VAS values derived from the SF-6D and EQ-5D are more difficult to explain. For the COPD, elderly female, osteoarthritis and hernia data sets, the VAS estimates obtained from the SF-6D were less than those obtained from the EQ-5D by 0.120, 0.119, 0.135, and 0.047. The differences were unlikely to have arisen owing to the SF-6D classification systematically describing the same health state as worse than the EQ-5D. A more obvious reason for the discrepancies was the source of the valuations. The values for the instruments were elicited from two very different samples of respondents. The Sheffield sample was younger and had only half

the proportion of responders over 65 (i.e. 13%) compared to the MVH sample (i.e. 24%). Age accounted for an 0.073 increase in SF-6D VAS health state valuation by the over 65's. Adjusting for age using this estimate, however, only accounts for a 0.01 point difference (i.e. this has been estimated as the 0.073 increase pro rata to the extra 11% of the sample who were over 65s).

A potentially more important source for the discrepancy could be the lower proportion of manual workers, and the higher number of people finishing education over 19 years of age in the Sheffield sample. To examine the likely impact of this on the derived EQ-5D VAS score, it was re-estimated using the algorithm estimated for the top educational group. For the COPD, elderly female and osteoarthritis groups this had the effect of reducing the difference by only 0.033, 0.025 and 0.036 respectively. The educational group accounts for a proportion of the difference but the discrepancy remains in excess of 0.1 (on the 0 to 1.0 scale). The difference between the SF-6D EQ-5D VAS values for hernia repair was reversed (i.e. from -0.048 to +0.02).

A final cause of the discrepancy could be the 'contextual effect'. The ratings of health states by VAS are known to be influenced by the seriousness of the other health states being valued (see Chapter 5). Loomes et. al. (1994) found that health states were assigned lower VAS values in a 'nice' group of states compared to those valued in 'nastier' groups. The health states defined by the SF-6D were on average substantially better than those defined by the EQ-5D, and hence were being rated in a 'nicer' (or less nasty) context in the Sheffield survey than in the MVH survey. The study by Loomes and colleagues found this contextual effect resulted in a difference of 0.06, 0.13 and 0.20 for three 'core' health states, and therefore this has the potential to be the main explanation for the differences between SF-6D and EQ-5D values in COPD, elderly female and osteoarthritis. However, this does not explain the result in the hernia group.

### 9.2.5 A CUA of alternative treatments of inguinal hernia[3]

The relative advantage of laparoscopic repair was significant at six weeks for the dimensions of pain, vitality, social functioning and physical functioning. These results would seem to reflect a faster recovery following the new procedure in comparison with open repair (Lawrence et al., 1995). This relative advantage diminished between six weeks and six months. Between baseline and six months the differences between patient groups in the SF-36 dimensions of pain, physical functioning and social functioning were in the same direction, though they were no longer significant at the 5% level.

These results indicate a marginal health benefit in favour of laparoscopic repair over the initial period, but was also the more expensive technique. It is therefore not possible to compare the alternative procedures in terms of their relative efficiency (see Chapter 4). The new preference-based algorithm can be used to translate these SF-36 data into a QALY gain in order to conduct a CUA.

The SG and VAS values at each assessment have been derived from the SF-36 data. These indicate a similar picture to the SF-36 dimensions, with a significant difference in favour of laparoscopic repair over the first six weeks, but a non-significant difference over 6 months at the 5% level (Table 9.18). The SG values have been translated into a QALY gain by estimating the area under the curve. This has been based on an assumption that the difference at six months is sustained for five years. The resultant QALY difference of 0.108 from laparoscopic repair has an 80% confidence interval (CI) of between 0.005 to 0.212. (The 80% CI was chosen rather than the conventional 95% CI in order to reflect better the degree of uncertainty likely to be acceptable in practical decision-making.) Lawrence and colleagues estimated the marginal cost of the laparoscopic procedures over and above the open procedure to be £582, with an 80% confidence interval of £434 to £730. This results in a central estimate for the marginal cost per QALY estimate of £5,389. The large ranges around the estimated marginal QALY gain and marginal cost indicate considerable uncertainty in this figure. Taking

---

[3] The authors of the study kindly agreed to provide the raw data required for the cost utility analysis undertaken in this chapter (Lawrence et al, 1995).

the extreme ends of the 80% confidence interval of each to reflect best and worst scenario indicates a potential range of £2,047 to £146,000.

The range in cost per QALY gained from laparoscopic surgery is very wide. The lower end of the range would compare favourably with published cost per QALY estimates for other interventions and the upper end would not (e.g. Williams, 1985; Maynard 1991). Comparisons of cost per QALY across studies and programmes should be undertaken with caution for a variety of reasons (Drummond et al., 1993). However, it is clear that the estimate is too uncertain to be able to assess the cost-effectiveness of investing in laparoscopic repair for inguinal hernia. A larger sample size would be required in order to reduce the extent of this uncertainty and the CUA would be further improved by a longer term follow up of the patients (as recommended by the authors Lawrence et al. (1995)).

## 9.3 Discussion and conclusion

The derivation of the SF-6D and the further reduction brought about by the modelling has resulted in some loss in the sensitivity of the SF-36, particularly in terms of responsiveness to health change. This loss may have been less if the modelling had been based on a larger valuation survey. The loss is also partly a result of the scoring algorithm for deriving the single value, which pools the changes across dimensions. The apparent reduction in responsiveness may reflect the strength of people's preferences for the overall change and not simply those changes that occur for one or two of the dimensions.

Despite the reduction in sensitivity, there is evidence to suggest that the SF-6D values have retained some of the advantages of the SF-36 over the EQ-5D in terms of descriptive validity at the milder end of the spectrum of illness. It was found, for example, that the SF-6D values were able to detect perceived health changes in COPD patients that were missed by the EQ-5D. There were too few studies, however, to be conclusive about the extent and generalisability of any advantage. The evidence on

empirical validity against stated preferences was also inconclusive since there was only VAS data.

To examine whether there are likely to be any practical implications for predicting patient choice or the ranking of interventions in terms of cost per QALY from using the SF-6D rather than the EQ-5D, mean health state values have been calculated for the five data sets used so far in this chapter. The values generated by these measures were found to rank the five samples in the same order i.e. the general population has the highest values, followed by hernia repair, elderly female, COPD, and osteoarthritis of the knee (Table 9.18). The SF-6D would not change the predicted choice between health states.

The size of the health state values and the intervals between the mean health state values of the samples, however, were very different. As would be expected, the intervals between the mean health state values of the samples were lowest for SG. What is more interesting was the finding that the intervals differ for the two sets of mean health state VAS values. These differences in the mean health state values would result in different size of the QALY gain from alternative interventions. This has important implications for predicting patient choice, evaluating the cost-effectiveness between alternatives for the same patient groups by CUA, and for making cross-programme comparisons. The best method of examining the practical importance of these potential differences would be to estimate cost per QALYs gained using EQ-5D and SG-6D on a number of data sets collected prospectively in randomised clinical trials. Unfortunately for the research in this thesis, there was only one published study collecting SF-36, EQ-5D and cost data as part of the trial at the time of writing. Furthermore, the EQ-5D data was incomplete and it was not possible to calculate QALYs gained using this measure.

The application of the SG algorithm to the trial of treatments for inguinal hernia demonstrated how SF-36 and cost results can be transformed into information suitable for assessing the cost-effectiveness of health care interventions. The primary purpose of the research reported in this thesis has been achieved, but the advantages of the SF-6D over existing preference-based measures of health is unproven owing to the absence of evidence.

**Table 9.1: Individual level models with background characteristics of respondents**

| | 1) Standard Gamble | | 2) VAS | |
|---|---|---|---|---|
| | B | T | B | T |
| PH2 | -.028 | -1.2 | -.053 | -2.3* |
| 3 | -.034 | -1.5 | -.098 | 4.1*** |
| 4 | -.075 | -3.0** | -.145 | 5.1*** |
| 5 | -.098 | -3.0** | -.155 | 4.3*** |
| 6 | -.104 | -3.3*** | -.258 | 7.4*** |
| R2 | .003 | .1 | -.123 | 3.7*** |
| S2 | .010 | .3 | .007 | .2 |
| 3 | -.037 | -1.1 | -.011 | .3 |
| 4 | -.051 | -1.4 | -.060 | 1.5 |
| 5 | .054 | 1.2 | -.189 | 3.9*** |
| Pain 2 | .013 | .5 | -.054 | 1.9 |
| 3 | .003 | .1 | -.011 | .4 |
| 4 | .004 | -.2 | -.104 | 4.1*** |
| 5 | -.131 | -3.9*** | -.121 | 3.1** |
| 6 | -.199 | -4.3*** | -.078 | 1.7 |
| M2 | -.014 | -.7 | -.078 | 3.6*** |
| 3 | -.001 | -.3 | -.156 | 5.1*** |
| 4 | .001 | .2 | -.155 | 4.6*** |
| 5 | -.258 | -5.3*** | -.036 | .7 |
| V2 | -.050 | -2.1* | -.002 | .1 |
| 3 | -.072 | -2.6* | -.030 | 1.0 |
| 4 | -.067 | -1.9 | -.029 | .8 |
| 5 | -.041 | -1.1 | -.102 | 2.4* |
| Age 45-65 | -.000 | -.0 | .026 | 1.7 |
| 65 and over | -.005 | -.3 | .076 | 3.5*** |
| Chronic illness | .005 | .5 | .012 | 1.7 |
| General Health | -.003 | .6 | -.002 | .4 |
| Sex | .003 | .3 | -.006 | .8 |
| Constant | .989 | 25.0*** | .903 | 32.3*** |
| df | 998 | | 1328 | |
| adj. $R^2$ | .340 | | .559 | |

**Table 9.2: Reliability in COPD patients who said their health had not changed over two six months periods**

| | Correlation | Mean difference | SD | n | 95% CI |
|---|---|---|---|---|---|
| a) Between initial assessment and six months | | | | | |
| **SF-6D** | | | | | |
| VAS | 0.47** | .000 | .127 | 49 | -.037, .036 |
| SG | 0.30* | .001 | .111 | 49 | -.031, .033 |
| **EQ** | | | | | |
| VAS | .56*** | -.009 | .156 | 41 | -.058.040 |
| TTO | .55*** | -.003 | .229 | 41 | -.075,.069 |
| b) Between six and twelve month assessments | | | | | |
| **SF-6D** | | | | | |
| VAS | .67*** | 0.15 | .113 | 35 | -.023,.054 |
| SG | .64*** | -.003 | .069 | 35 | -.027, 0.20 |
| **EQ** | | | | | |
| VAS | .67*** | -.025 | .166 | 36 | -.081.031 |
| TTO | .65*** | -.047 | .274 | 36 | -.139, .046 |

**Table 9.3: Reliability in an elderly (>75) female population who said their health had not changed over a six month period.**

| | n | Correlation | Mean difference | SD | 95% CI |
|---|---|---|---|---|---|
| **SF-6D** | | | | | |
| VAS | 66 | .70*** | .008 | .136 | -.025, .041 |
| SG | 66 | .67*** | .006 | .075 | -.013, .024 |
| **EQ** | | | | | |
| VAS | 66 | .66*** | .013 | .143 | -.023,.48 |
| TTO | 66 | .63*** | .014 | .207 | -.037,.065 |

**Table 9.4: Descriptive validity - general population sample**

| | SF-6D VAS | SF-6D SG |
|---|---|---|
| **Age (years):** | | |
| 16-24 | .764 | .942 |
| 25-34 | .750 | .938 |
| 35-44 | .696 | .922 |
| 45-54 | .677 | .914 |
| 55-64 | .636 | .895 |
| 65-74 | .569*** | .867 |
| **Socioeconomic class:** | | |
| I | .729 | .935 |
| II | .737 | .937 |
| III non-manual | .701 | .703 |
| III manual | .716 | .725 |
| IV | .682 | .914 |
| V | .625** | .890* |
| **GP consultation in previous 2 weeks:** | | |
| Yes | .605 | .880 |
| No | .722*** | .930*** |
| **O/P attendance in previous 3 months:** | | |
| Yes | .597 | .876 |
| No | .716*** | .928*** |
| **I/P admission in last year:** | | |
| Yes | .621 | .881 |
| No | .709*** | .925*** |

* P< 0.05, **P<0.01 and ***P<0.001 by Kruskal Wallis one-way ANOVA

**Table 9.5: Descriptive validity - Elderly (>75) female population**

| Health Indicator | n | SF-6D VAS | SF-6D SG | EQ-5D VAS | EQ-5D TTO |
|---|---|---|---|---|---|
| GP in last 2 weeks | | | | | |
| Yes | 93 | .433 | .796 | .572 | .622 |
| No | 239 | .514*** | .841*** | .621* | .548* |
| O/P attendance in last 3 mths | | | | | |
| Yes | 103 | .460 | .811 | .571 | .550 |
| No | 226 | .510* | .839* | .628* | .631* |
| A&E attendance in last 3 mths | | | | | |
| Yes | 44 | .419 | .812 | .548 | .507 |
| No | 279 | .506** | .780** | .621 | .622 |
| Hospital stay in last year | | | | | |
| Yes | 49 | .419 | .788 | .528 | .497 |
| No | 274 | .507** | .837*** | .622** | .620** |
| Any long standing illness | | | | | |
| Yes | 180 | .418 | .795 | .535 | .506 |
| No | 104 | .570*** | .865*** | .688*** | .704*** |
| OPCS Disability classification | | | | | |
| 0 | 27 | .661 | .915 | .787 | .840 |
| 1 | 58 | .613 | .888 | .712 | .743 |
| 2 | 44 | .561 | .871 | .672 | .694 |
| 3 | 30 | .498 | .829 | .679 | .707 |
| 4 | 48 | .464 | .812 | .578 | .562 |
| 5 | 36 | 0432 | .806 | .574 | .560 |
| 6 | 34 | .398 | .779 | .506 | .459 |
| 7 | 30 | .361 | .766 | .429 | .355 |
| 8 | 17 | .328 | .739 | .412 | .346 |
| 9 | 6 | .340*** | .730*** | .170*** | .070*** |

**Table 9.6: Descriptive validity - COPD patients**

| Health Indicator | n | SF-6D VAS | SF-6D SG | EQ-5D VAS | EQ-5D TTO |
|---|---|---|---|---|---|
| Breathlessness | | | | | |
| Severe | 26 | .340 | .767 | .472 | .464 |
| Less Severe | 75 | .476*** | .842*** | .671*** | .699*** |
| Exercise Tolerance (6 min walk test) | | | | | |
| End VAS | | | | | |
| >65 | 21 | .396 | .822 | .623 | .646 |
| ≤65 | 21 | .375 | .795 | .554 | .554 |
| Distance | | | | | |
| ≤302 | 18 | .339 | .788 | .640 | .668 |
| >302 | 25 | .425* | .824* | .512** | .494** |
| $FEV_1$ % Predicted | | | | | |
| ≤.41 | 51 | .366 | .785 | .528 | .519 |
| >.41 | 50 | .369 | .774 | .558 | .547 |
| Hospital Admission | | | | | |
| Yes | 26 | .302 | .743 | .472 | .448 |
| No | 78 | .394** | .797** | .550** | .550* |
| Comorbidity | | | | | |
| Yes | 59 | .330 | .752 | .497 | .468 |
| No | 65 | .394* | .800** | .562 | .567 |

**Table 9.7: Cross-tabulation of responses to EQ-6D and SF-6D by dimension in the general population sample**

| | | Physical | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| EQ1 | 1 | 579 | 363 | 308 | 99 | 22 | 47 |
| Mobility | 2 | | | 3 | 3 | 8 | 28 |
| | 3 | | | | | 1 | |
| EQ2 | 1 | 581 | 365 | 312 | 102 | 35 | 64 |
| | 2 | | | 1 | | | 10 |
| Self care | 3 | | | | 1 | | 1 |

| | | Role | |
|---|---|---|---|
| | | 1 | 2 |
| EQ3 | 1 | 877 | 566 |
| Main activity | 2 | 2 | 49 |

| | | Social | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| EQ4 | 1 | 982 | 284 | 96 | 64 | 14 |
| Leisure | 2 | 13 | 11 | 12 | 27 | 16 |

| | | Pain | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| EQ5 | 1 | 561 | 263 | 99 | 45 | 8 | 1 |
| | 2 | 27 | 90 | 150 | 208 | 37 | 4 |
| Pain/Distress | 3 | | | | 2 | 19 | 9 |

| | | Mental Health | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| EQ6 | 1 | 851 | 267 | 56 | 15 | 12 |
| Anxiety/ | 2 | 26 | 75 | 82 | 67 | 32 |
| Depression | | | | | | |

**Table 9.8: Patients at the ceiling of the EQ classification by SF-6D dimension in the general population sample**

| | n | Age (mean) | Sex %F | % GP visit in last 2 weeks | % O/P Attendence in last 3 mth | % I/P Admission in last yr |
|---|---|---|---|---|---|---|
| **Physical (where EQ1=1 and EQ2=1** | | | | | | |
| **1** | 578 | 33 | 50.2 | 15.2 | 9.7 | 8.9 |
| **2** | 361 | 42 | 50.0 | 16.7 | 11.7 | 8.0 |
| **3** | 308 | 47 | 65.6 | 21.0 | 13.1 | 8.8 |
| **4** | 99 | 52 | 68.8 | 33.0 | 26.0 | 21.9 |
| **5** | 22 | 57 | 50.0 | 41.0 | 41.0 | 31.8 |
| **6** | 46 | 47*** | 52.2*** | 41.3*** | 22.3*** | 13.0** |
| **Role (EQ3=1)** | | | | | | |
| **1** | 907 | 39 | 49.8 | 12.6 | 10.2 | 8.5 |
| **2** | 644 | 43 | 63.7*** | 27.9 | 18.4 | 12.1 |
| **Social (EQ4=1)** | | | | | | |
| **1** | 974 | 41 | 49.5 | 13.1 | 10.3 | 8.1 |
| **2** | 282 | 40 | 64.4 | 26.2 | 17.1 | 15.1 |
| **3** | 96 | 41 | 65.6 | 30.2 | 28.7 | 13.7 |
| **4** | 63 | 39 | 71.9 | 42.9 | 16.1 | 7.8 |
| **5** | 14 | 33 | 69.2*** | 28.6*** | 0.0*** | 14.3* |
| **Pain (EQ5=1)** | | | | | | |
| **1** | 559 | 37 | 48.3 | 11.0 | 6.4 | 7.0 |
| **2** | 263 | 38 | 61.8 | 21.0 | 9.7 | 10.0 |
| **3** | 97 | 35 | 66.7 | 23.0 | 13.3 | 14.1 |
| **4** | 45 | 38 | 60.0 | 40.0 | 15.6 | 2.2 |
| **5** | 8 | 32 | 75.0 | 62.5 | 12.5 | 12.5 |
| **6** | 1 | 19 | 0.0*** | 0.0*** | 0.0*** | 100.0 |
| **Mental (EQ6=1)** | | | | | | |
| **1** | 845 | 41 | 47.4 | 14.8 | 11.3 | 9.5 |
| **2** | 264 | 42 | 61.1 | 21.6 | 17.9 | 12.0 |
| **3** | 56 | 42 | 75.0 | 31.0 | 16.1 | 12.5 |
| **4** | 15 | 42 | 80.0 | 20.0 | 6.7 | 0.0 |
| **5** | 12 | 49 | 58.3*** | 58.3*** | 17.0 | 8.3 |

*p<0.05, **p<0.01, ***p<0.001 by Mantel-haenszel test for linear association

**Table 9.9: Cross-tabulation of responses of EQ-6D and the VAS SF-6D model (where different from SF-6D)**

| | | Social | | |
|---|---|---|---|---|
| | | 1 & 2 | 3 | 4 & 5 |
| EQ4 | 1 | 1266 | 96 | 78 |
| Leisure | 2 | 24 | 12 | 43 |

| | | Pain | | | |
|---|---|---|---|---|---|
| | | 1 | 2 & 3 | 4 | 5 & 6 |
| EQ5 | 1 | 561 | 362 | 45 | 9 |
| | 2 | 27 | 240 | 208 | 41 |
| Pain/Distress | 3 | | | 2 | 28 |

| | | Mental Health | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 & 5 |
| EQ6 | 1 | 851 | 267 | 56 | 27 |
| Anxiety/Depression | 2 | 26 | 75 | 82 | 99 |

**Table 9.10: Cross-tabulation of responses to EQ-6D and the SG SF-6D model (where different from SF-6D)**

| | | Physical | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 & 5 | 6 |
| EQ1 | 1 | 579 | 363 | 308 | 121 | 47 |
| Mobility | 2 | | | 3 | 11 | 28 |
| | 3 | | | | 1 | |

| | | Social | | |
|---|---|---|---|---|
| | | 1 & 2 | 3 | 4 & 5 |
| EQ4 | 1 | 1266 | 96 | 78 |
| Main Activity | 2 | 24 | 12 | 43 |

| | | Pain | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 & 3 | 4 | 5 | 6 |
| EQ5 | 1 | 561 | 362 | 45 | 8 | 1 |
| Leisure | 2 | 27 | 240 | 208 | 37 | 4 |
| | 3 | | | 2 | 19 | 9 |

| | | Mental Health | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 & 4 | 5 |
| EQ6 | 1 | 851 | 267 | 71 | 12 |
| Pain/ Distress | 2 | 26 | 75 | 149 | 32 |

**Table 9.11: Patients at the ceiling of the EQ classification by VAS SF-6D model level (where different from SF-6D)**

| | n | Age (mean) | Sex % F | % GP visits | % 0/P attendance in last 2 weeks | % 1/P admission in last year |
|---|---|---|---|---|---|---|
| **Social (EQ4=1)** | | | | | | |
| **S1 & 2** | 1256 | 41 | 52.8 | 16.1 | 11.8 | 9.7 |
| **3** | 96 | 41 | 65.6 | 30.2 | 29.0 | 13.7 |
| **4 & 5** | 77 | 38 | 71.4*** | 40.3*** | 13.2 | 9.0 |
| **Pain (EQ5=1)** | | | | | | |
| **1** | 559 | 37 | 48.3 | 10.9 | 6.4 | 7.0 |
| **2 & 3** | 362 | 37 | 63.2 | 21.4 | 10.6 | 11.1 |
| **4** | 45 | 38 | 60.0 | 40.0 | 11.1 | 2.2 |
| **5 & 6** | 9 | 31 | 66.7** | 55.6*** | 15.6* | 22.3 |
| **Mental (EQ6=1)** | | | | | | |
| **1** | 845 | 41 | 47.4 | 14.8 | 11.3 | 9.5 |
| **2** | 264 | 42 | 61.1 | 21.6 | 17.9 | 12.0 |
| **3** | 56 | 42 | 75.0 | 31.0 | 16.1 | 12.5 |
| **4 & 5** | 27 | 45 | 70.4*** | 37.1*** | 11.1 | 3.7* |

**Table 9.12: Patients recording no problem on the EQ classification by SG SF-6D model**

| | n | Age (mean) | Sex % F | % GP visits | % 0/P attendance in last 2 weeks | % 1/P admission in last year |
|---|---|---|---|---|---|---|
| **Physical** | | | | | | |
| **1** | 433 | 32 | 49.7 | 13.8 | | 8.8 |
| **2** | 177 | 41 | 47.5 | 13.7 | 5.1 | 6.4 |
| **3** | 127 | 46 | 66.1 | 15.8 | 7.2 | 7.1 |
| **4 & 5** | 22 | 54 | 68.2 | 18.2 | 18.2 | 22.3 |
| **6** | 18 | 38*** | 27.8 | 33.3* | 0.0 | 0.0 |
| **Social** | | | | | | |
| **1 & 2** | 745 | 38 | 51.1 | 13.0 | 6.4 | 7.8 |
| **3** | 25 | 34 | 64.0 | 36.0 | 20.8 | 4.0 |
| **4 & 5** | 18 | 34 | 66.7 | 47.1*** | 5.6 | 5.6 |
| **Pain** | | | | | | |
| **1** | 475 | 37 | 45.7 | 9.7 | 5.6 | 6.4 |
| **2 & 3** | 271 | 37 | 62.4 | 18.6 | 7.9 | 10.4 |
| **4** | 34 | 38 | 52.9 | 35.3 | 14.7 | 3.0 |
| **5** | 8 | 32 | 75.0 | 42.5 | 12.5 | 12.5 |
| **6** | 1 | 19** | 0.0 | 0.0*** | 0.0 | 100.0* |
| **Mental** | | | | | | |
| **1** | 598 | 37 | 47.3 | 12.1 | 6.3 | 7.4 |
| **2** | 137 | 37 | 63.5 | 18.4 | 9.8 | 8.8 |
| **3 & 4** | 45 | 39 | 84.4 | 24.7 | 6.7 | 6.7 |
| **5** | 5 | 33 | 40.0** | 60.0*** | 0.0 | 0.0 |
| **Vitality** | | | | | | |
| **1 & 2** | 672 | 38 | 49.9 | 12.4 | 7.4 | 7.3 |
| **3, 4 & 5** | 111 | 34 | 67.5*** | 27.3*** | 3.6 | 9.1 |

**Table 9.13: SF -6D VAS and EQ-5D VAS compared to patients' own VAS rating**

| Estimated VAS | n | Correlation | Mean Difference (Derived VAS minus own **) | SD | 95% CI |
|---|---|---|---|---|---|
| COPD | | | | | |
| SF-VAS | 121 | .43 | -.141 | .169 | -.110,-.171 |
| EQ-VAS | 120 | .55 | -.021 | .174 | -.010,-.053 |
| Elderly Females | | | | | |
| SF-VAS | 318 | .54 | -.183*** | .177 | -.163,-.202 |
| EQ-VAS | 312 | .49 | -.064*** | .197 | -.086,-.042 |
| OA knee | | | | | |
| SF-VAS | 210 | .56 | -.283*** | .177 | -.259,-.307 |
| EQ-VAS | 208 | .51 | -.148*** | .208 | -.176,-.120 |
| Hernia | | | | | |
| SF-VAS | 132 | .54 | -.104*** | .149 | -.129,-.078 |
| EQ-VAS | 133 | .40 | -.057*** | .140 | -.081,-.033 |

**Table 9.14: Correlations of the difference between the derived VAS values and the patients' own rating and the derived value for EQ and SF by patient group**

| Patient Group | EQ-VAS | SF-VAS |
|---|---|---|
| **COPD** | .52*** | .52*** |
| **Elderly female** | .43*** | .40** |
| **Hernia** | .26** | .56*** |
| **OA** | .47*** | .02 |

**Table 9.15: Responsiveness - mean differences by health change in COPD patients**

| | Worse | | | Same | | | Better | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean difference | (SD) | n | Mean difference | (SD) | n | Mean difference | (SD) | n | ANOVA P |
| **SF-6D** | | | | | | | | | | |
| VAS | -.035 | (.118) | 33 | .000 | (.109) | 49 | .046 | (.095) | 20 | .0125 |
| SG | -.020 | (.094) | 33 | .000 | (.111) | 49 | .021 | (.074) | 20 | .1406 |
| **EQ** | | | | | | | | | | |
| VAS | -.051 | (.121) | 30 | .001 | (.110) | 42 | -.058 | (.126) | 22 | |
| TTO | -.093 | (.224) | 30 | .000 | (.229) | 41 | -.082 | (.185) | 22 | |

**Table 9.16: Correlation coefficients between SF-6D and EQ by condition**

| **1) General Population** | **SF-6D** | **EQ-6D** | |
|---|---|---|---|
| | SG | VAS | |
| SF-6D VAS | .90 | | |
| SF-6D SG | | | |
| | | | |
| **2) COPD** | **SF-6D SG** | **EQ-5D VAS** | **EQ-5D TTO** |
| SF-6D VAS | .81 | .52 | .50 |
| SF-6D SG | | .53 | .54 |
| EQ-5D VAS | | | .99 |
| | | | |
| **3) Elderly Female** | **SF-6D SG** | **EQ-5D VAS** | **EQ-5D TTO** |
| SF-6D VAS | .86 | .68 | .65 |
| SF-6D SG | | .69 | .68 |
| EQ-5D VAS | | | .99 |
| EQ-5D TTO | | | |
| | | | |
| **4) OA knee** | **SF-6D SG** | **EQ-5D VAS** | **EQ-5D TTO** |
| SF-6D VAS | .84 | .57 | .54 |
| SF-6D SG | | .66 | .65 |
| EQ-5D VAS | | | .99 |
| EQ-5D TTO | | | |
| | | | |
| **5) Hernia** | **SF-6D SG** | **EQ-5D VAS** | **EQ-5D TTO** |
| SF-6D VAS | .89 | .70 | .67 |
| SF-6D SG | | .74 | .74 |
| EQ-5D VAS | | | .99 |
| EQ-5D TTO | | | |
| | | | |
| **6) Overall** | **SF-68 SG** | **EQ-5D VAS** | **EQ-5D TTO** |
| SF-6D VAS | .86 | .62 | .59 |
| SF-6D SG | | .66 | .65 |
| EQ-5D VAS | | | .99 |
| EQ-5D TTO | | | |

**Table 9.17: Comparisons of SF-6D and EQ-5D health state values for five samples**

| | SF-6D | | | | EQ-5D | | | |
|---|---|---|---|---|---|---|---|---|
| Sample | VAS mean | diff.* | SG mean | diff. | VAS mean | diff. | TTO mean | diff. |
| General pop. | .699 | | .92 | | .837 | | - | |
| | | .01 | | .002 | | .09 | | |
| Hernia repair | .689 | | .918 | | .739 | 8 | .775 | |
| | | .196 | | .088 | | .13 | | .172 |
| Elderly female | .493 | | .83 | | .609 | | .603 | |
| | | .127 | | .051 | | .07 | | .081 |
| COPD | .366 | | .779 | | .533 | 6 | .522 | |
| | | .052 | | .034 | | .08 | | .155 |
| OA knee | .314 | | .745 | | .446 | 7 | .367 | |

* This is the difference between the mean health state value of this data set and the previous one.

**Table 9.18: SF-36 scores and SF-6D VAS and SG values in patients with inguinal hernia treated by open (O) and laparascopic (L) surgical techniques.**

| SF-36 | | Baseline | 10 days | 6 weeks | Overall gain up to 6 weeks | Difference 95% CI ( ) | 3 mths | 6 mths | Overall gain up to 6 months | 95% CI of difference |
|---|---|---|---|---|---|---|---|---|---|---|
| Physical functioning | | | | | | | | | | |
| | O | 87 | 62 | 88 | - 11.5 | | 93 | 94 | 2.4 | |
| | L | 80 | 66 | 88 | - 4.5 | | 92 | 94 | 7.2 | |
| | | | | | | 7.0(13.0, 1.1)* | | | | 4.8 (-11.1,-1.5) |
| Role (P) | | | | | | | | | | |
| | O | 81 | 14 | 55 | - 44.1 | | 89 | 92 | -13.1* | |
| | L | 66 | 30 | 65 | -18.1 | | 86 | 96 | -3.8 | |
| | | | | | | 26.0(13.3,38.6)*** | | | | 9.8 (18.1, 4)* |
| Role (E) | | | | | | | | | | |
| | O | 90 | 76 | 83 | -8.5 | | 92 | 94 | 1.5 | |
| | L | 88 | 69 | 77 | -15.4 | | 85 | 97 | -5.5 | |
| | | | | | | -6.9 (2.9,-16.7) | | | | -7.0 (11.1, -1.5) |
| Social | | | | | | | | | | |
| | O | 82 | 51 | 75 | -17.5 | | 84 | 83 | -2.1 | |
| | L | 76 | 59 | 80 | -6.9 | | 83 | 85 | 1.3 | |
| | | | | | | 10.6 (-16.8,4.3)*** | | | | 3.4 (9.3, -3.2) |
| Pain | | | | | | | | | | |
| | O | 74 | 45 | 76 | -13.7 | | 90 | 85 | 7.3 | |
| | L | 66 | 60 | 81 | 2.5 | | 84 | 91 | 14.2 | |
| | | | | | | 16.2(-24.1,-8.2)*** | | | | 6.9 (16.6, -2.8) |
| Mental Health | | | | | | | | | | |
| | O | 80 | 81 | 83 | 1.5 | | 84 | 85 | 5.5 | |
| | L | 77 | 80 | 83 | 3.3 | | 81 | 83 | 4.1 | |
| | | | | | | 4.8 (-5.4,-1.8) | | | | -1.4 (4.0, -6.8) |
| Vitality | | | | | | | | | | |
| | O | 69 | 54 | 66 | -7.9** | | 73 | 76 | 4.3 | |
| | L | 65 | 60 | 71 | -.4 | | 71 | 73 | 3.9 | |
| | | | | | | 8.3 (-12.3,2.7)** | | | | -0.4 (6.1, -6.9) |
| GHP | | | | | | | | | | |
| | O | 80 | 76 | 78 | -2.8 | | 83 | 84 | 2.3 | |
| | L | 75 | 71 | 77 | -1.4 | | 76 | 79 | 1.6 | |
| | | | | | | 4.2 (-5.4,-2.5) | | | | -0.7 (4.4,-5.8) |
| **SF-6D** | | | | | | | | | | |
| VAS | O | .708 | .521 | .707 | -.092* | | .805 | .798 | .047 | |
| | L | 664 | .585 | 735 | -.014 | | .770 | .816 | .080 | |
| | | | | | | 0.78 (122. 034)* | | | | .033 (090,-023) |
| SG | | | | | | | | | | |
| | O | 927 | 857 | 931 | -.034* | | .963 | .956 | .015 | |
| | L | 908 | 889 | 936 | 001 | | .947 | .963 | .032 | |
| | | | | | | .033 ( 055.- 014)* | | | | .017 (.044,-010) |

# Chapter 10

# Discussion and conclusion

The core of the thesis is the development of a preference-based measure of health from the SF-36. It is primarily a methodological thesis and the main contributions are to the methods of benefit valuation in the economic evaluation of health care interventions. There are, however, some insights of more general interest for applied economics.

The chapter begins by examining the contributions of the research reported in this thesis to economic evaluation in health care methodology, and then discusses the more general implications for health benefit valuation and the elicitation of preferences. This is followed by a discussion of future research in this area.

## 10.1 Contributions of the research

### 10.1.1 Economic evaluation in health care

Extending the coverage of economic evaluation in health care

The SF-36 is potentially a rich source of data for economic evaluation, since it has become one of the most widely used measures of general health in clinical trials being conducted in the UK, the rest of Europe and North America. The algorithms for deriving health state values from SF-36 data are simple to apply and the author has written a short computer programme on SPSS for Windows. These algorithms can be used to transform a set of SF-36 data, largely unsuitable for use in economic evaluation in its current form, to undertake a CUA (as was demonstrated in the application to the results of the hernia repair trial). This has the potential of extending the application of CUA to the results of clinical trials which would otherwise only be suitable for a cost-consequences analysis. The significance of this new capability can be gauged from the large number of requests received by the author from researchers wishing to use it.

The use of these algorithms should be tempered by the preliminary nature of the research on which they are based. There are many limitations with this study, largely as a result of the resource constraints. For example, the SF-6D needs to be critically reviewed and refined in the light of the modelling work, and the judgements tested against patient values. The valuation survey can be criticised for its comparatively small scale (at least compared to the more recent surveys to value the EQ-5D and HUI-III), and the unrepresentativeness of the sample of respondents. Furthermore, the algorithms were based on a reduced version of the SF-6D brought about by the merging of dimension levels to eliminate inconsistencies due to collinearity in the dimension levels. (The section on future research considers improvements that can be made to overcome these limitations).

For the moment, the preliminary algorithms resulting from this work provide the only way of undertaking a CUA using SF-36 (and cost) data, but users must be made aware of these limitations.

A new preference-based measure of health

One solution to the problems with using general profile measures of health such as the SF-36 would be to use an existing preference measure alongside it. As the review in Chapter 3 found, four of the five existing preference-based measures are brief and easy to use self-administered questionnaires and would add little burden to data collection in a clinical trial (i.e. the Health Measurement Questionnaire for the Rosser disability/distress classification, the HUI-I to III, the EQ-5D, and the 15D). The size of the contribution of this research to economic evaluation depends on whether the new preference-based measure of health is an improvement on existing measures.

The strength of the original SF-36 lay in the descriptive validity of its dimensions and their sensitivity. There is evidence of its greater sensitivity compared to the Rosser, and more importantly the EQ-5D, at detecting milder conditions and responding to health

changes. The superiority of the new preference-based measure partly rests on whether the adaptation of the SF-36 into the SF-6D and the further simplifications brought about by the modelling has substantially reduced the sensitivity of the original instrument. The comparison with EQ-5D was important because this has been identified as the best of the existing five preference-based measures of health.

The SF-6D values have retained some of the advantages of the SF-36 over the EQ-5D in terms of descriptive validity at the milder end of the spectrum of illness. There were too few studies, however, to be conclusive about the extent and generalisability of any advantage. The advantage with the EQ-5D is that it benefits from the results of a large, well-conducted survey of a representative sample of the general adult population of the UK. The valuation data set was considerably larger and of better quality than that used to value SF-6D (being based on preferences elicited by interview). On the other hand, some economists would prefer the SF-6D since it has algorithms based on preferences obtained by SG rather than TTO. As illustrated in the review of elicitation techniques (in Chapter 3), the theoretical position of TTO compared to SG remains an area of contention in the health economics literature and therefore this would not be regarded as an advantage by many economists. Finally, the evidence on empirical validity against stated preferences was also inconclusive since there was only VAS data.

### 10.1.2 Health Benefit Valuation

<u>A new approach to obtaining preferences for health status measures</u>

This research was the first attempt to derive a preference-based value from a profile measure of health status. Multi-dimensional scales have been used to value health and other characteristics of goods in applied economics areas before, but these scales were designed for the purpose. What makes the SF-36 and other measures of health status more difficult for valuation is their size and the lack of ordinality between many of the items within dimensions as well as between dimensions. The research reported in this thesis developed an approach for dealing with the problems arising from this and this

would be useful to others contemplating a similar exercise with another measure of health status.

Adaptation of the SG question

This was only the second time that SG has been directly administered to respondents to elicit their preferences over a health state classification, and the first time that the results have been used to estimate a model for health state values. The valuation survey confirmed the findings of the MVH pilot survey, that the self-completed version of the SG questionnaire can be used to obtain consistent health state values (at least for gambles with a fatal outcome for treatment failure). The responses of the patients, many of whom were quite elderly, did show higher levels of inconsistency. There was no formal qualitative evidence, but remarks made during the sessions indicated many of the patients had difficulties understanding the task. An interview administered method with the aid of props would be advised for this group.

As has become common practice in surveys to elicit preferences for health, the respondents were asked to consider a scale of probabilities rather than an open ended question. A disadvantage with this approach is the restrictions it places on the respondents choice of categories. An important innovation in this survey was to add four more response categories between 0.95 and 1.00 to the list of probabilities of success in order to improve the sensitivity of scale for the milder states of the SF-36. The view taken in designing the valuation survey was that the scale in the original version of the question may not have been sufficiently sensitive at the upper end of rates of success. It would seem highly unlikely, for example, that someone would choose an operation with a one in twenty risk of death to cure a chronic medical condition in health state 211111. This view was found to be correct, with the four additional categories being chosen as the point of indifference in over a third of the responses (437/1243).

There was an option for respondents to choose their own for chances of success for probabilities in excess of 0.95 in the original question and therefore it could be argued these additions were unnecessary. However, there is likely to be an inertia on the part of respondents to 'opt out' of the scale. This hypothesis was supported by the low numbers who chose to do so: just 6.6% (82) selected their own value between 0.99 and 1.0. Future research using SG should examine more critically the response choices available in the question.

The alternative solution to using these comparatively high levels of risk of death for valuing health states with mild or moderate health problems is to change the treatment failure reference state. However, this was shown in this study and elsewhere to lead to respondent confusion and generate values inconsistent with the axioms of expected utility theory.

Use of statistical techniques

The decision to use multivariate statistical techniques was based on a review of the literature and discussions with statisticians experienced in the use of such techniques. It was encouraging to find from this research that these techniques were successful in valuing a classification larger than the EQ-5D, with just 1293 observations compared with 34,298 from the MVH survey. Simple additive models were able to achieve good levels of fit for subjective stated preference data, and though there was evidence of heterogeneity in the SG individual model, it was found to be robust in a re-run of the model on two random samples of the data set. A large number of observations may not be necessary for multivariate valuation work. These results should be encouraging to other researchers considering using the same approach who do not have the comparatively high level of resources that were available to the MVH group. In retrospect, however, the selection of health states for valuation in the survey proved to be more of a problem. Considerable multi-collinearity was found between the dimensions, and this may have been responsible for the inconsistencies and made it difficult to distinguish interactions from the main effect. Future valuations of health

state classifications, particularly larger ones such as SF-6D, must be more systematic in the selection of states.

An important problem encountered in the modelling work was the hierarchical structure of the data set from respondent variation since this invalidates the conventional OLS assumption of independence of the errors. A fixed effects adjustment for the variation between respondents was found to substantially improve the fit of the model. This method permits differences between respondents in terms of the intercept. A more complex multi-level modelling approach was applied to stated preference data for the first time. Using the statistical package MLn, it was possible to explore more complex error structures involving dimension levels. These random effects were found to significantly improve the fit of the model and reduce the size of the standard errors on the coefficients. An interesting finding was the positive association between the size of the variance terms and dimension level. However, these more sophisticated error structures did not change the size of the beta coefficients, and hence had no implication for the algorithm. This finding has important implications for analyses of these types of data. At the request of Paul Dolan of the York MVH group, the author undertook the same multi-level analysis on the EQ-5D using their TTO data. The result was the same. Improvements were achieved in model efficiency, but the results suggested that the simpler model which limited between respondent variation to the constant term did not bias the coefficients. The improvements in efficiency of both the SF-6D and EQ-5D models were small, and did not justify the added complexity.

### 10.1.3 Relationship between VAS and SG

Models using VAS health state values to estimate SG values achieved a good fit, and passed the diagnostic tests for model specification. However, the estimated parameters in the models were different to those published elsewhere, which suggest this is an unreliable relationship. This would cast doubt on using VAS in place of direct valuation by SG.

The instability in parameter values may be due to the findings at individual level, which provided further evidence against the relative risk attitude explanation suggested by Dyer and Sarrin (1982) and Torrance et al. (1995) as the sole explanation for the differences. They have argued that the only difference between VAS and SG is a persons (constant) attitude to risk, and VAS has been regarded as measure of value under certainty. There was evidence consistent with RRA, but the models also contained a positive non-zero intercept, lending support to the Gambling effect hypothesis. Furthermore, there are competing explanations for the concave power function. All models fitted the data poorly at the individual level. This evidence supports the findings of other published studies and suggests there other important explanations.

### 10.1.4 VAS as a technique for eliciting strength of preferences

There has been considerable scepticism among economists as to whether the VAS can be regarded as a cardinal measure of preferences at all. VAS valuations are subject to 'spreading' and 'context' effects and interviews with respondents in other studies have suggested there is no strength of preference intention. A comparison of the coefficients of the health dimensions and their levels estimated from valuation data collected by VAS and SG supports this argument. The most important dimension in the VAS model was physical functioning, whereas severe pain and mental health were more important for the SG model. These results suggest that VAS seems to be a measure of health in terms of concepts such as physical fitness, rather than a reflection of preferences. This finding combined with evidence from elsewhere indicates that VAS should not be used to derive preferences (though it may continue to have a role as a preliminary task prior to SG or some other choice based technique).

### 10.1.5 Implications for expected utility theory (EUT)

The comparison of SG health state values obtained from a single gamble involving death as the worst reference state, with those calculated from two gambles where one values the same state with a non-fatal worst reference health state, has implications for EUT. It suggests a significant departure from the predictions of EUT and may reflect

important violations in the axioms of EUT. This would support a growing body of evidence of such violations. The divergence may be explained by an aversion to gambling effect or simply an inability to adjust to the change in reference gambles. The latter reflects a natural limit to peoples cognitive abilities and this has important implications for economic theory, as well as health state valuations. This also raises more profound concerns about the questionnaire approach to preference elicitation. The implication for empirical work is that the worst reference state should not be changed.

## 10.2 Future Research

The research reported in this thesis has a number of shortcomings. These were mainly the result of the limited resources available to the study. The following research agenda would address these shortcomings and thereby improve the algorithms for deriving a preference-based measure for health from SF-36 data.

Revisions to the SF-6D health state classification

The SF-6D was derived by a multi-disciplinary team and involved many subjective judgements. Further work is required to test these judgements against the views of patients. The health classification could also be improved from the knowledge, experience and insights of other researchers, and in particular the original developers of the SF-36.

Valuation surveys

There were a number of shortcomings to be overcome in a future valuation survey. The design of a future survey should therefore incorporate the following:

- A larger and more representative sample of the constituency of interest. (For informing resource allocation decisions a sample of the adult general population is usually recommended). A larger sample would result in more reliable and significant estimated parameter values. It would also be important for ensuring the values are more representative and this could be important since the MVH main survey found

the background characteristics of respondents had a significant effect on peoples valuations.

- Undertake a systematic selection of health states designed to address the multi-collinearity between dimensions of the SF-6D. The ideal solution is to sample health states using a statistical factorial design, but the selection would have to be limited to those states which are creditable and found to occur in practice.

- Use an interviewer administered version of SG with more categories at the top end of the scale or a procedure which makes it easier to select your own value.

There has been interest from researchers in other countries in using the results of this research. Comparisons of the results of surveys to value the EQ by VAS undertaken in different countries suggests the differences may be small. The extent of variation in health state value choice based techniques of elicitation, such as SG is not known. There would be a case for undertaking valuation surveys in other countries.

Modelling

The opportunity to examine alternative specifications was limited by the number of observations and the multi-collinearity between dimensions. A larger and better designed survey would permit a more complex model specification to be examined. The larger data set would be more difficult to analyse using a fixed-effects adjustment and therefore a random-effects component should be used to allow for respondent variation. This would also represent a more efficient use of the data. The results presented in this thesis suggest it would only be necessary to allow for variations in the intercept.

Comparison with EQ-5D

The systematic review found the EQ-5D to be currently the best preference-based measure of health (though the HUI-III would be another contender when the new algorithms is published). It will be important to compare the SF-6D (or its successor) to the EQ-5D in terms of the criteria of practicality, reliability, descriptive validity, validity

of its values and empirical validity. This should include a series of comparative studies to examine the descriptive validity of both measures, covering a range of conditions, interventions, age groups and settings. The research should include an assessment of content validity using patient interviews and evaluating their construct validity. The ultimate test, of course, is whether the values reflect preferences. The empirical validity of the measure should be compared in terms of revealed preferences (where possible), stated preferences, and hypothesised preferences.

## 10.3 Conclusion

This thesis has been concerned with the adaptation of a health status questionnaire into an instrument for use in cost-utility analysis. The SF-36 has been revised into a multi-dimensional classification amenable to valuation. This classification has been valued by the use of a self-administered version of the SG elicitation technique and the application of multivariate modelling techniques. The research has been successful in estimating a set of preference-based algorithms for valuing the SF-36. The adoption of the new algorithms will extend the application of economic evaluation in health care. Furthermore, they provide an alternative to existing preference-based measures for estimating QALYs and may prove to be more suitable in some circumstances, particularly for milder conditions. There is considerable scope, however, for further research to improve the classification, the survey and the technique of modelling health state values.

## References


Aaronson, N.K., Acquadro, C., Alonso, J., Apolone, G., Bucquet, D., Bullinger, M., Bungay, K., Fukuhara, S., Gandek, B., Keller, S., & et al. (1992). International Quality of Life Assessment (IQOLA) Project. *Quality in Life Research, 1*, 349-351.

Abdalla, M., Russell, I. (1995). Tariffs for the Euroqol health states based on modelling individual VAS and TTO data of the York Survey In : MVH Group *Final Report on the Modelling of Valuation Tariffs* Centre for Health Economics, University of York, UK.

Adomowicz W., Louiviere J., Williams M. (1994). Combining Revealed and Stated Preference methods for valuing environmental amenities. *Journal of Environmental Economics and Management;* 26: 271-292.

Allais, M. (1953). Le Comportement de l'Homme Rationnal devant le Risque, Critique des Postulants et Axiomees de l'Ecole Americaine. *Econometrica*; 21, 503-546.

Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School. In M. Allais and O Hagen (eds). Expected Utility Hypothesis and the Allais Paradox, Dordrecht: Reidel.

Anderson, G.M. (1982). A comment on the index of well-being. *Medical Care, 20,* 513-515.

Anderson, J.P., Bush, J.W., & Berry, C.C. (1986). Classifying function for health outcome and quality-of-life evaluation. Self- versus interviewer modes. *Medical Care, 24,* 454-469.

Anderson, J.P., Bush, J.W., & Berry, C.C. (1988). Internal Consistency Analysis: a method for studying the accuracy of function assessment for health outcome and quality of life evaluation. *Journal Clinical Epidemiology, 41,* 127-137.

Anderson, J.P., Kaplan, R.M., Berry, C.C., Bush, J.W., & Rumbaut, R.G. (1989). Interday reliability of function assessment for a health-status measure - the quality of well-being scale. *Medical Care, 27,* 1076-1084.

Anderson, J.P., Kaplan, R.M., & Schneiderman, L.J. (1994). Effects of offering advance directives on quality adjusted life expectancy and psychological well-being among ill adults. *Journal Clinical Epidemiology, 47,* 761-772.

Anderson, R.T., Aaronson, N.K. and Wilkin, D. (1993) Critical review of the international assessments of health-related quality of life. *Quaity in Life Research.* **2**, 369-395.

Andresen, E.M., Patrick, D.L., Carter, W.B. and Malmgren, J.A. (1995) Comparing the performance of health status measures for healthy older adults. *Journal of the American Geriatriatric Society.* **43**, 1030-1034.

Apajasalo, M., Sintonen, H., Holmberg, C., Sinkkonen, J., Aalberg, V., Pihko, H., Siimes, M.A., Kaitila, I., Makela, A., Rantakari, K., Anttila, R. and Rautonen, J. (1996) Quality-of-life in early adolescence - a 16-dimensional health-related measure (16d). *Quality of Life Research* **5**, 205-211.

Appleby L. & Stammer S. (1987). Individual choice under uncertainty: A Review of Experimental Evidence, past and present. In survey in economics of uncertainty eds. Key J, Lambert P. Basil Blackwell, Oxford.

Arrow, K.J. (1951). Social Choice and Individual values. New York: Wiley.

Arrow, K.J. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*; 53, 941-73.

Arrow, K.J., Lind, R.C. (1970). Uncertainty and the evaluation of public investment decisions. *American Economic Review;* 60:364-378

Ashby, J., O'Hanlon, M., & Buxton, M.J. (1994). The time trade-off technique: how do the valuations of breast cancer patients compare to those of other groups? *Quality in Life Research, 3,* 257-265.

Backhouse, M., Backhouse R., & Edey, S.A. (1992). Economic Evaluation Bibliography. *Health Economics, 1* (supplement).

Badia X, Fernandez E, Segura A (1995) Influence of socio-demographic and health status variables on evaluation of health states in a spanish population *European Journal Public Health*,5:87-93.

Bakker, C.H., Rutten van Molken, M., van Doorslaer, E., Bennett, K. and van der Linden, S. (1993) Health related utility measurement in rheumatology: an introduction. *Patient. Educ. Couns.* **20**, 145-152.

Balaban, D.J., Sagi, P.C., Goldfarb, N.I. and Nettler, S. (1986) Weights for scoring the quality of well-being instrument among rheumatoid arthritics. A comparison to general population weights. *Medical. Care* **24**, 973-980.

Barr, R.D., Feeny, D., Furlong, W., Weitzman, S., & Torrance, G.W. (1995). A preference-based approach to health-related quality-of-life for children with cancer. *International Journal Of Paediatric Haematology/Oncology, 2,* 305-315.

Barr, R.D., Furlong, W., Dawson, S., Whitton, A.C., Strautmanis, I., Pai, M., Feeny, D., & Torrance, G.W. (1993). An assessment of global health status in survivors of acute lymphoblastic leukaemia in childhood. *American Journal Pediatric Hematoligical Oncology, 15,* 284-290.

Barr, R.D., Pai, M.K.R., Weitzman, S., Feeny, D., Furlong, W., Rosenbaum, P., & Torrance, G.W. (1994). A multi-attribute approach to health status measurement and clinical management -illustrated by an application to brain tumours in childhood. *International Journal of Oncolology, 4,* 639-648.

Bass, E.B., Steinberg, E.P., Pitt, H.A. et al (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.

Bateman, I., Munro, A., Rhodes, B., Starmer, C., Sugden, R. (1997) Does part-whole bias exist? - an experimental investigation. *Economic Journal*; 107(441): 322-332

Bates, J. (1988). Papers on Stated Preference Methods In Transport Research. *Journal of Transport Economics and Policy*, 22: 59-70.

Becker, G.S. (1962). Investment in human capital: a theoretical analysis. *Journal of Political Economy, 70,* 9-49.

Bell, D. (1982). Regret in Decision Making under Uncertainty. *Operations Research*; 30, 961-981,.

Bell, D. (1985). Disappointment in Decision Making under Uncertainty. *Operations Research*; 33,
1-27.

Benson, T.J.R. (1978). Classification of disability and distress by ward nurses: a reliability study. *International Journal of Epidemeology;* 7: 359-361.

Bergner, M., Bobbit, R.A., Carter, W.B., Gibson, B.S. (1981). The Sukness Impact Profile: development and final revision of a health status measure. *Medical Care*; 18: 787-805.

Bernoulli, D. (1938). Specimen Theoriad Novae de Mensura Sortis. Commentarii Academiae Scientiarum Impariales Petrapolitane; 5, 175-92. (Translated by L. Somer: Exposition of a New Theory on the Measurement of Risk, *Econometrica*; 22, 23-36 1954).

Berwick, D.M., Murphy, J.A, Goldman, P.A. et al (1991). Performance of a five item mental health screening test. *Medical care*; 29 (2): 169-76.

Billson A L, Walker D A (1994) Assessment of health status in survivors of cancer Archives of disease in Childhood; 70:200-204

Bindman, A.B., Keane, D., Luirie, N. (1990). Measuring Health changes among severely ill patients - the floor phenomenon. *Medical Care*; 28: 1442-52.

Bjork, S (1991) Euroqol Conference Proceedings. Swedish Health Economics Institue Discussion paper 1

Bleichrodt, H. (1995). QALYs and HYEs (healthy year equivalents): under what conditions are they equivalent? *Journal of Health Economics, 14,* 17-37.

Boadway, R.W., Bruce, N. (1984). Welfare Economics. Blackwell, Oxford.

Boland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* I: 307-10.

Bombardier, C., & Raboud, J. (1991). A comparison of health-related quality-of-life measures for rheumatoid-arthritis research. *Controlled Clinical Trials, 12,* S 243-S 256.

Bombardier, C., Ware, J., Russell, I., Larson, M.G., Chalmers, A. and Leighton Read, J. (1986) Auranofin therapy and quality of life in patients with rheumatoid arthritis. *The American Journal of Medicine* **81**, 565-578.

Bombadier C, Wolfson AD, Sinclair AJ, McGreer A (1982). Comparison of three measurement methodologies in the evaluation of functional status Index In: Deber R and

Thompson G (eds) *Choices in Health Care: Decision Making and Evaluation of Effectiveness*, Toronto: University of Toronto.

Bosch, J.L., Hunink, M.G.M. (1996). The Relationship between Descriptive and Valuational Quality-of-life Measures in Patients with Intermittent Claudication. *Medical Decision Making*; 16: 217-225.

Bowe, T.R. (1995). Measuring Patient Preferences: Rating scale versus Standard Gamble. *MedicalDecision Making*; 15(3): 283-285.

Bowling, A. (1991). Measuring health: a review of quality of life and measurement scales. Milton Keynes: Open University Press.

Boyle, M.H., Furlong, W., Feeny, D., Torrance, G.W., & Hatcher, J. (1995). Reliability of the Health Utilities Index--Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. *Quality Life Research, 4,* 249-257.

Boyle, M.H., Torrance, G.W., Sinclair, J.C., & Horwood, S.P. (1983). Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *NewEngland Journal of Medicine, 308,* 1330-1337.

Boyle, M.H., Torrance, G.W. (1984). Developing multi-attribute health indexes. *Medical Care, 22,* 1045-1057.

Bradley, M. (1988). Econometric issues in stated preference analysis. Journal of Transport Economics and Policy; 17, 269-288.

Bradlyn, A.S., Harris, C.V., Warner, J.E., Ritchey, A.K., & Zaboy, K. (1993). An investigation of the validity of the Quality of Well-Being Scale with paediatric oncology patients. *Health Psychology, 12,* 246-250.

Brazier, J.E. (1995). The SF-36 Health Survey and its use in Pharmaco-economic Evaluation *Pharmaco-economics* 7 (5): 403-415.

Brazier, J.E., Deverill, M., Harper, R., Booth, A. (1997) A Review of the use of health status questionnaires in economic evaluation. Final report to the NHS Executive HTA Programme.

Brazier, J.E., Dixon, S. (1995). The use of condition specific outcome measures in economic appraisal. *Health Economics, 4,* 255-264.

Brazier, J., Harper, R., Jones, N.M.B. et al (1992) Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *British Medical JournalL,* 305: 160-4.

Brazier, J., Jones, N., & Kind, P. (1993). Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. *Quality in Life Research, 2,* 169-180.

Brazier, J., Snaith,M.,Munro, J. (1996b) Measuring health outcomes in people with osteoarthritis of the knee. Report submitted to the NHS Executive (Trent).

Brazier, J.E., Usherwood, T.P., Harper, R., Jones, N.M.B., Thomas, K. (1994). Deriving a single index measure for health from the SF-36 - Interim Report for the Department of Health, Medical Care Research Unit, University of Sheffield.

Brazier, J., Walters, S.J., Nicholl, J.P., & Kohler, B. (1996a). Using the SF-36 and Euroqol on an elderly population. *Quality in Life Research, 5,* 195-204.

Brook, R.H., Ware, J.E., Robers, W.R. et al (1983). Does free care improve adults' health? Results from a randomised controlled trial. *New England Journal of Medicine*; 309: 1426-1434.

Brooks, R.G. (1995). Health Status Measurement: A perspective on change. McMillan, Basingstoke and London.

Brooks, R.G., Jendteg S, Lindgren B, Persson U, & Bjork 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.

Bryan, S., Parkin, D., & Donaldson, C. (1991). Chiropody and the qaly - a case-study in assigning categories of disability and distress to patients. *Health Policy, 18,* 169-185.

Buckingham, K. (1993). A note on hye (healthy years equivalent). *Journal of Health Economics, 12,* 301-309.

Buckingham, K. (1995). Economics, health and health economics - HYEs versus QALYs - a response. *Journal of Health Economics, 14,* 397-398.

Bush, J.W., Anderson, J.P., Kaplan, R.M. and Blischke, W.R. (1982) Counter-intuitive preferences in health-related quality-of-life measurement. *Medical Care* **20**, 516-525.

Buxton, M., Acheson, R., et al (1985) Costs and Benefits of the Heart Transplant programmes at Harefield and Papworth hospitals, DHSS Office of the Chief Scientist Research Report No. 12, London:HMSO.

Cadman, D., Goldsmith, C. (1986). Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. *Journal of Chronic Disease*; 39(8): 643-651.

Cadman, D., Goldsmith, C. and Bashim, P. (1984) Values, preferences and decisions in the care of children with developmental disabilities. *Developmental and Behavioral Pediatrics* **5**, 60-64.

Cairns, J.A. (1994). Valuing future benefits. *Health Economics, 3,* 221-229.

Cairns, J., Johnston, K. (1992). Assessing the severity of depressive illness. *Journal of Clinical Psychology*. July 1992, Vol. 48(4): 455-462.

Cairns, J., Johnston, K., McKenzie, L. (1991) Developing QALYs from condition-specific outcome measures. HERU DP 14/91. University of Aberdeen

Calfas, K.J., Kaplan, R.M. and Ingram, R.E. (1992) One-year evaluation of cognitive-behavioral intervention in osteoarthritis. *Arthritis Care Research.* **5**, 202-209.

Caperna, J. and Mathews, W.C. (1996) Estimating health-related quality-of-life (hr-qol) among persons with hiv-infection using the euroqol instrument - do the euroqol health dimensions explain self-rated global health. *Journal Of Investigative Medicine* **44**, A 155

Carr-Hill, R.A. (1989). Assumptions of the QALY procedure. *Social Science and Medicine*, *29*, 469-477.

Carr-Hill, R.A. (1991) A good measure for Eurohealth? *Health Service Journal*. **101**, 24-25.

Carr-Hill R.A. (1992). A second opinion: Health-related quality of life measurement -- Euro style. *Health Policy*, *20*, 321-328.

Carr-Hill, R.A., & Morris, J. (1991). Current practice in obtaining the q in QALYs - a cautionary note. *British Medical Journal*, *303*, 699-701.

Carter, W., Bobbitt, R., Bergner, M., Gibson, B (1976). Validation of an interval scaling: the Sickness Impact Profile. *Health Service Research*; 11: 516-28.

Cassidy, H.J. (1981) *Using Econometrics: A Beginners guide*.

Cattin, P., Wittink, D. (1982). Commercial use of conjoint analysis: a survey. *Journal of Marketing*, 14, 21-33.

Chan, C.L.H. and Villar, R.N. (1996) Obesity and quality-of-life after primary hip-arthroplasty. *Journal Of Bone And Joint Surgery-British Volume* **78B**, 78-81.

Coast, J. (1992). Reprocessing data to form QALYs. *British Medical Journal*, *305*, 87-90.

Coast, J. (1993a). Developing the qaly concept - exploring the problems of data-acquisition. *PharmacoEconomics*, *4*, 240-246.

Cole, R.P., Shakespeare, V., Shakespeare, P. and Hobby, J.A. (1994) Measuring outcome in low-priority plastic surgery patients using Quality of Life indices. *British Journal of Plastic Surgery* **47**, 117-121.

Cook, J., Richardson, J., & Street, A. (1994). A cost-utility analysis of treatment options for gallstone disease - methodological issues and results. *Health Economics*, *3*, 157-168.

Culyer, A.J. (1971a). The nature of the commodity health care and its efficient allocation. *Oxford Economic Papers, 24,* 189-211.

Culyer, A.J. (1971b), Medical Care and the Economics of Giving. *Economica*; 151, 295-303.

Culyer, A.J. (1971c). Merit Goods and the Welfare Economics of Coercion. *Public Finance*; 26, 546-71.

Culyer, A.J. (1978), Measuring Health: Lessons for Ontario, Toronto, University of Toronto Press.

Culyer, A.J. (1989a), Commodities, Characteristics of Commodities, Characteristics of People, Utilities and Quality of Life, in Baldwin, S., Godfrey, C., Propper, C. (eds). The Quality of Life: Perspectives and Policies, London, Routledge.

Culyer, A.J. (1989b) The normative economics of health care finance and provision *Oxford Review of Economic Policy*, 5(1): 34-58.

Culyer, A.J., Simpson, H. (1980) Externality models and health: A Ruckbick over the last ten years. *Economic Record*; 56: 222-30.

Culyer, A.J., & Wagstaff, A. (1993). QALYs versus HYEs. *Journal of Health Economics*, *12*, 311-323.

Culyer, A.J. Lavers, R., Williams, A. (1971). Social Indicators: Health, *Social Trends*; 2, 31-42.

Culyer, A.J., van Doorslaer, E., & Wagstaff, A. (1992). Utilisation as a measure of equity by Mooney, Hall, Donaldson & Gerard. *Journal of Health Economics*, *11*, 93-98.

Currim, I.S., Sarin, R.K. (1984). A comparative evaluation of multi-attribute consumer preference models. Management Science. 30(5): 543-561.

Darnell, A. (1992). Decision-making under uncertainty. In Maloney J, *What's new in Economics*? Manchester University Press, Manchester and New York; 1-39.

Davies, O.L. (1956). *The designs and analysis of industrial experiments*. Oliver and Boyd.

Deaton, A., Muellbauer, J. (1980). *Economics and Consumer behaviour*. Cambridge University Press, Cambridge.

Debreu, G., (1959). *Theory of value: An axiomatic study of Economic Equilibrium*. New York: Wiley.

de Groot, J., de Groot, W., Kamphuis, M., Vos, P.F., Berend, K. and Blankestijn, P.J. (1994) [Little difference in quality of life of dialysis patients in Utrecht and Willemstad] Kwaliteit van leven van dialysepatienten in Utrecht en Willemstad weinig verschillend. *Ned. Tijdschr. Geneeskd.* **138**, 862-866.

Deyo, R.A., Inui, T.S. (1984). Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research*; 19(3): 275.

Dirksen, S.R. (1995) Search for meaning in long-term cancer survivors. *Journal of Advanced Nursing.* **21**, 628-633

Dolan, P. (1994) Search for a critical-appraisal of Euroqol - a response by the Euroqol group to Gafni and Birch. *Health Policy* **28**, 67-69.

Dolan, P. (1995). Modelling valuation for Euroqol health states. Paper presented to an HESG meeting, University of Aberdeen.

Dolan, P., Gudex, C., Kind, P. and Williams, A. (1995) A social tariff for Euroqol: Results from a UK general population survey, Centre for Health Economics Discussion Paper 138, University of York.

Dolan, P., Gudex, C., Kind, P. and Williams, A. (1996) Valuing health states: a comparison of methods. *Journal of Health Economics*, 2, 209-232.

Dolan, P., Sutton, M. (1997) Mapping visual analogue scores onto time trade-off and stanard gamble utilities. *Social Science and Medicine*, forthcoming

Donaldson, C. (1993). Theory and practice of willingness to pay for health care. HERU DP 01/93, University of Aberdeen.

Donaldson, C. (1995) Willingness to pay for publicly-provided health care. Thesis prsented for the degree of doctor of philosophy at the University of Aberdeen.

Donaldson, C., Atkinson, A., Bond, J., & Wright, K. (1988a). Should QALYs be programme-specific? *J Health Economics*, *7*, 239-257.

Donaldson, C., Atkinson, A., Bond, J. and Wright, K. (1988b) QALYS and long-term care for elderly people in the UK: scales for assessment of quality of life. *Age and Ageing* **17**, 379-387.

Donaldson, C., & Gerard, K. (1993). *Economics of Health Care Financing: the Visible Hand.* Macmillan, London.

Donaldson,C., Hundley, V., Mapp, T. (1995a) Willingness to pay: A new method for measuring patients' preferences? HERU Discussion Paper, University of Aberdeen

Donaldson, C., Shackley, P. (1997) Does "process utility" exist? A case study of willingness to pay for laporoscopic cholecystectomy. *Social Science and medicine*, 44(5), 285-294.

Donaldson, C., Shackley, P., Abdalla, M., Miedzybrozka, Z. (1995b) Willingness to pay for antenatal carrier screening for cystic fibrosis. *Health Economics*, 4, 439.

Donaldson, C., Thomas, R., Torgeson. (1997) Validity of open-ended and payment scale approaches to eliciting willingness to pay. *Applied Economics*, 29, 79-84.

Drewett, R.F., Minns, R.J., Sibly, T.F. (1992). Measuring outcome of total knee replacement using quality of life indices. *Ann R Coll Surg Engl*; 74, 286-289.

Drummond, M.F. (1992). Cost-effectiveness guidelines for reimbursement of pharmaceuticals: is economic evaluation ready for its enhanced status? *Health Economics*, 1: 85-92.

Drummond, M.F., Davies, L. (1991). Economic Analysis alongside clinical trials: Revisiting the methodological issues. *International Journal of Technology Assessment In Health Care*; 7(4):561-573).

Drummond, M.F., Ludbrook, A., Lowson, K.V., & Steele, A. (1986). *Studies in Economic Appraisal in Health Care: Volume 2.* Oxford University Press, Oxford.

Drummond, M.F., Stoddart, G.L., Torrance, G.W. (1987). Methods for the economic evaluation of health care programmes. Oxford: *Oxford Medical Publications*.

Drummond, M., Torrance, G., & Mason, J. (1993). Cost-effectiveness league tables: more harm than good? *Social Science and Medicine*, *37*, 33-40.

Dupuit, J. (1844) On the measurement of utility of public works. Translated by RH Barback in International Economic papers, 2 (1952): 83-110 from "De la Measure de l'utilite des Travaux Public" Annales des Ponts et Chausses, 2nd series Vol 8.

Dyer, J.S. Sarin, R.K. (1982). Relative risk aversion. *Management Science*, 28:875-886.

Ellsberg, D. (1961). Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*; 75, 643-669.

Ellwood, P.M. (1988). Outcomes management: a technology of patient experience. *New England Journal of Medicine*; 318: 1549-56.

Elvik, R. (1995). The validity of using health state indexes in measuring the consequences of traffic injury for public-health. *Social Science and Medicine*; 40, 1385-1398.

Eraker, S.A., Baker, S., & Miyamoto, J.M. (1984). Parameter estimates for a quality adjusted life year utility model. *Clinical Research, 32,* A294.

Erickson, P., Kendall, E.A., Anderson, J.P. and Kaplan, R.M. (1989) Using composite health status measures to assess the nation's health. *Medicine Care* **27**, S66-76.

Essink-Bot, M.L., Bonsel, G.J., Van Der Mass, P.J. (1990). Valuation of health states by the general public: Feasibility of a standardised measurement procedure. *Social Science and Medicine*; 31, 1201-1206.

Essink-Bot, M.L., Stouthard, M.E. and Bonsel, G.J. (1993) Generalizability of valuations on health states collected with the EuroQolc-questionnaire. *Health Economics;* **2**, 237-246.

Essink-Bot, M.L., Vanroyen, L., Krabbe, P., Bonsel, G.J., Rutten, F.F.H. (1995). The impact of migraine on health-status. *Headache*; 35, 200-206.

Euroqol group. (1990). Euroqol - a new facility for the measurement of health-related quality-of-life. *Health Policy, 16,* 199-208.

Euroqol group (1991) Not a quick fix (response to Carr-Hill). *Health Services Journal.* **101**, 29

Euroqol group (1992) Euroqol -- a reply and reminder. *Health Policy* **20**, 329-332.

Evans, R.G., & Wolfson, A.D. (1980). *Faith, hope and charity: health care in the utility function.* Department of Economics, University of British Columbia and Department of Health Administration, University of Toronto, unpublished paper.

Feeny, D.H., Torrance, G.W. (1989). Incorporating utility based quality of life assessment measures in clinical trials. *Medical Care*; 27(3): S190-204.

Feeny, D., Furlong, W., Barr, R.D., Torrance, G.W., Rosenbaum, P., & Weitzman, S. (1992). A comprehensive multi attribute system for classifying the health status of survivors of childhood cancer. *Journal of Clinical Oncoologyl, 10,* 923-928.

Feeny, D., Furlong, W., Boyle, M., & Torrance, G.W. (1995). Multi-attribute health status classification systems. Health Utilities Index. *PharmacoEconomics, 7,* 490-502.

Feeny, D., Leiper, A., Barr, R.D., Furlong, W., Torrance, G.W., Rosenbaum, P., & Weitzman, S. (1993). The comprehensive assessment of health status in survivors of childhood cancer: application to high-risk acute lymphoblastic leukaemia. *British Journal of Cancer, 67,* 1047-1052.

Feeny, D., Labelle, R. Torrance, G. W. (1990) Intergrating economic evaluations and quality of life assessments. In: *Quality of life assessments in clinical trials*. ed Spilker, B. New York, Raven Press.

Feeny, D., Torrance, G.W., Goldsmith, C., Furlong, W., & Boyle, M. (1994). *A multi-attribute approach to population health status.* Hamilton, Ontario: CHEPA McMaster University.

Feldstein,M.S. (1963) Economic analysis, operational research, and the National Health Service, *Oxford Economic Papers*, 15:19-31.

Fitzpatrick, R.(1991) Surveys of patient satisfaction:!-Imporatant general considerations. *British Medical Journal*, 303:887-889.

Fitzpatrick, R., Zeibland, S., Jenkinson, C., Mowat, A. (1993) A comparison of the sensitivity to change of several Health Status Measures in Rheumatoid Arthritis. *Journal of Rheumatology*, 20:429-36.

Fowkes, T., Wardman, M. (1988) The design of stated preference travel choice experiments. The *Journal of Transport Economics and Policy*, 22:27-44.

Friedman, L. S. (1984) *Microeconomic Policy Analysis*. New York, McGraw-Hill.

Froberg, D.G., & Kane, R.L. (1989a). Methodology for measuring health-state preferences--I: Measurement strategies. *Journal Clinical Epidemiology*, *42*, 345-354.

Froberg, D.G., & Kane, R.L. (1989b). Methodology for measuring health-state preferences--II: Scaling methods. *Journal of Clinical Epidemiology*, *42*, 459-471.

Fryback, D.G., Dasbach, E.D., Klein, R., Klein, B.E.K., Martin, P.A., Dorn, N. and Peterson, K. (1992) Health assessment by SF-36, Quality of Well-Being index and time trade-offs: predicting one measure from another. *Medical Decision Making*. **12**, 348P

Fryback, D.G., Dasbach, E.J., Klein, R., Klein, B.E., Dorn, N., Peterson, K. and Martin, P.A. (1993) The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Medical Decision making*. **13**, 89-102.

Fuchs, V.R. (1966) The contribution of Health Services to the American Economy. *Milbank Memorial Fund Quarterly*, 44:65-101.

Furlong, W., Torrance, G.W. and Feeny, D. (1995) Properties of Health Utilities Index: preliminary evidence. *Quality of Life Newsletter* 3-10.

Gafni, A. (1994). The standard gamble method: what is being measured and how it is interpreted. *Health Services Research*, *29*, 207-224

Gafni, A., & Birch, S. (1991). Equity considerations in utility-based measures of health outcomes in economic appraisals - an adjustment algorithm. *Journal of Health Economics*, *10*, 329-342.

Gafni, A., & Birch, S. (1993). Searching for a common currency - critical-appraisal of the scientific basis underlying European harmonisation of the measurement of health related quality-of-life (Euroqol(c)). *Health Policy*, *23*, 219-228.

Gafni, A., & Birch, S. (1995). Preferences for outcomes in economic evaluation: an economic approach to addressing economic problems. *Social Science and Medicine*, *40*, 767-776.

Gafni, A., Torrance, G.W. (1984) Risk attitude and time preference in health. *Management Science*, 30:440-451.

Gafni, A., & Zylak, C.J. (1990). Ionic versus non-ionic contrast media: a burden or a bargain? *Canadian Medical Association Journal, 143,* 475-478.

Gafni, A., Birch, S., & Mehrez, A. (1993). Economics, health and health economics - HYEs versus QALYs. *Journal of Health Economics, 12,* 325-339.

Ganiats, T.G., Humphrey, J.B., Taras, H.L., & Kaplan, R.M. (1991). Routine neonatal circumcision: a cost-utility analysis. *Medical Decision Making, 11,* 282-293.

Ganiats, T.G., Miller, C.J., & Kaplan, R.M. (1995). Comparing the quality-adjusted life-year output of 2 treatment arms in a randomised trial. *Medical Care, 33,* AS245-AS254.

Ganiats, T.G., Palinkas, L.A., & Kaplan, R.M. (1992). Comparison of Quality of Well-Being scale and Functional Status Index in patients with atrial fibrillation. *MedicalCare, 30,* 958-964.

Garratt, A.M., Ruta, D.A., Abdalla, M.I. et al (1993). The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS. *British Journal of Medicine*; 306: 1440-4.

Gater, R.A., Kind, P. and Gudex, C. (1995) Quality of life in liaison psychiatry. A comparison of patient and clinician assessment. *British Journal of Psychiatry* **166**, 515-520.

Office of Population Censuses and Surveys (1990) *General Household Survey 1988*, London: HMSO, 1990.

Gerard, K. (1992). Cost-utility in practice - a policy makers guide to the state-of-the- art. *Health Policy, 21,* 249-279.

Gilbert, A., Owen, N., Innes, J.M. and Sansom, L. (1993) Trial of an intervention to reduce chronic benzodiazepine use among residents of aged-care accommodation. *Ausralien and New Zealand Journal of Medicine.* **23**, 343-347.

Glasziou, P.P., Bromwich, S. and Simes, R.J. (1994) Quality of life six months after myocardial infarction treated with thrombolytic therapy. AUS-TASK Group. Australian arm of International tPA/SK Mortality Trial. *Medicine J. Aust.* **161**, 532-536.

Gold, M., Franks, P. and Erickson, P. (1996) Assessing the health of the nation: the predictive value of a preference based measure and self-rated health. *Medical. Care* **34**, 163-177.

Goldstein (1995) *Multilevel Statistical Methods* London: Edward Arnold, New York: Halstead Press.

Gravelle, H. (1995) Valuations of Euroqol health states: comments and suggestions.Paper presented at the ESRC/SHHD Workshop on Quality of Life, Edinburgh, unpublished.

Gravelle, H., Rees, R. (1991) *Microeconomics*. Longman, London.

Greene, W.H. (1993). *Econometric Analysis* (2nd edition). Macmillan, New York.

Grogono, A.W., & Woodgate, D.J. (1971). Index for measuring health. *Lancet*, 1024-1026.

Grossman, M. (1972) On the concept of health capital and the demand for health. *Journal of Political Economy*, 80:223-55

Gudex, C. (1986). QALYs and their use by the Health Service. Discussion Paper 20, Centre for Health Economics, University of York.

Gudex, C.M. (1995) Health-related quality of life in endstage renal failure. *Quality in Life Research.* **4**, 359-366.

Gudex, C. and Kind, P. (1988) The QALY toolkit. Centre for Health Economics Discussion Paper 93, University of York.

Gudex, C. and Kind, P. (1991) Chiropody and the qaly - a case-study in assigning categories and distress to patients. *Health Policy* **19**, 79-80.

Gudex, C., Kind, P., van Dalen, H., Durand M-A, Morris, J., & Williams, A. (1993). Comparing scaling methods for health state valuations: Rosser revisited. Centre for Health Economics Discussion Paper 107, University of York.

Gudex, C., Williams, A., Jourdan, M., Mason, R., Maynard, J., O'Flynn, R. and Rendall, M. (1990) Prioritising waiting lists. *Health Trends.* **22**, 103-108.

Gujarati, W.H. (1993) *Basic Econometrics*. McGraw-Hill Book Company, New York.

Guyatt, G. (1993). Measurement of health-related quality of life in chronic airflow limitation. *Monaldi Arch Chest Dis, 48,* 554-557.

Hall, J., Gerard, K., Salkeld, G., & Richardson, J. (1992). A cost utility analysis of mammography screening in Australia. *Social Science and Medicine, 34,* 993-1004.

Harper, R., Brazier, J.E.,Waterhouse, J.C., Walters, S., Jones, N., Howard, P. (1997) a comparison of outcome measures for patients with chronic obstructive pulmonary disease in an outpatient setting. Submitted to Thorax

Hey, J.D., Lambert, P.J. (1987) *Surveys in the economics of uncertainty. Oxford. Basil Blackwell.*

Hicks, J.R. (1939). The foundations of welfare economics. *Economic Journal, 49,* 696-710.

Hicks, J.R. (1941). The rehabilitation of consumer's surplus. *The Review of Economic Studies, 8,* 108-116.

Holbrook, T.L., Hoyt, D.B., Anderson, J.P., Hollingsworth-Fridlund, P. and Shackford, S.R. (1994) Functional limitation after major trauma: a more sensitive assessment using the Quality of Well-being scale--the trauma recovery pilot project. *Journal of Trauma.* **36**, 74-78.

Hollingworth, W., Mackenzie, R., Todd, C.J., & Dixon, A.K. (1995). Measuring changes in quality-of-life following magnetic-resonance- imaging of the knee - SF-36, Euroqol((c)) or Rosser index. *Quality of Life Research, 4,* 325-334.

Hornberger, J.C., Redelmeier, D.A. and Petersen, J. (1992) Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *Journal Clinical Epidemeology* **45**, 505-512.

Humphreys, W.V., Evans, F., Watkin, G. and Williams, T. (1995) Critical limb ischemia in patients over 80 years of age -options in a district general-hospital. *British Journal Of Surgery* **82**, 1361-1363.

Hunt, S.M., McEwen, J., McKenna, S.P. (1986). Measuring Health Status. Croom Helm, London.

Hurst, N.P., Jobanputra, P., Hunter, M., Lambert, M., Lochhead, A., Brown, H. (1994). Validity of Euroqol - a generic health status instrument - in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. *British Journal of Rheumatology*; 33, 655-662.

Hurst N A (1996) longitudinal study of patients with rheumatoid arthritis Unpublished presentation to a Clinical Users Group of the EQ-5D, York,

Jenkinson, C., Coulter, A., Wright, L (1993). SF-36 health survey questionnaire: normative data for adults of working age. *British Journal of Medicine*; 306: 1437-1440.

Jenkinson, C., Layte, R., Wright, L., Coulter, A. (1996). *The UK SF-36: an analysis and interpretation manual*. Health Services Research Unit, University of Oxford, Oxford.

Johannesson, M. (1994). QALYs, HYEs and individual preferences-a graphical illustration. *Social Science and Medicine, 39,* 1623-1632.

Johannesson, M. (1995a). Quality-adjusted life-years versus healthy-years equivalents - a comment. *Journal of Health Economics, 14,* 9-16.

Johannesson, M. (1995b). Qalys - a comment. *Journal Of Public Economics, 56,* 327-328.

Johannesson, M. (1995c). 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., Jonsson, B., Karlson, G. (1996) Outcome Measurement in economic evaluation. *Health Economics*, 5(4): 279-298.

Johannesson, M., Pliskin, J.S., & Weinstein, M.C. (1994). A note on QALYs, time trade-off, and discounting. *Medical Decision Making, 14,* 188-193.

Jones, P.W. Quality of life measurement for patients with diseases of the airways. *Thorax*; 1991, 46: 676 - 6892.

Jones-Lee, M.W. (1989). *The Economics of Safety and Physical Risk*. Basil Blackwell, Oxford.

Jones-Lee, M., Loomes, G., O'Reilly D., Phillips, P. (1993). *The value of preventing non-fatal road injuries: findings of a willingness-to-pay national sample survey.* Transport Research Laboratory.

Kahneman, D., Knetch, J. (1992). Valuing public goods: the purchase of moral satisfaction. *J. Environmental Economics & Management*; 22: 57-70.

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

Kaldor, N. (1939). Welfare propositions of economics and interpersonal comparisons of utility. *Economic Journal*, 49: 549-52.

Kallis, P., Unsworth White, J., Munsch, C., Gallivan, S., Smith, E.E., Parker, D.J., Pepper, J.R. and Treasure, T. (1993) Disability and distress following cardiac surgery in patients over 70 years of age. *Europeon Journal of Cardiothorac Surgery.* **7**, 306-311.

Kanabar et al (1995) Quality of life in survivors of childhood cancer after megatherapy with autologous bone marrow rescue. *Pediatric Hematology and Oncology*, 12:29-36

Kaplan, R.I. and Atkins, C.J. (1989) The well-year of life as a basis for patient decision-making. *Patient. Educ. Couns.* **13**, 281-295.

Kaplan, R.M. (1989). Health outcome models for policy analysis. *Health Psycholology*, 8. 723-735.

Kaplan, R.M. (1993a) Application of a general health policy model in the American health care crisis. *Journal of the Royal Society of Medicine* **86**, 277-281.

Kaplan, R.M. (1993b). Quality of life assessment for cost/utility studies in cancer. *Cancer Treat Rev, 19 Suppl A,* 85-96.

Kaplan, R.M. (1994a) Using quality-of-life information to set priorities in health-policy. *Social Indicators Research* **33**, 121-163.

Kaplan, R.M. (1994b) Value judgment in the oregon medicaid experiment. *Medical Care* **32**, 975-988.

Kaplan, R.M., & Anderson, J.P. (1988). A general health policy model: update and application. *Health Services Research, 23,* 203-235.

Kaplan, R.M., & Anderson, J.P. (1990). The general health policy model: an integrated approach. In Anonymous, *Quality of life assessments in clinical trials.* New York: Raven Press Ltd.

Kaplan, R.M., Anderson, J.P., & Ganiats, T.G. (1993). S. Walker & R. Rosser (Eds.), *Quality of life assessments: key issues in the 1990s.* (pp. 65-93). Kluwer Academic Publishers.

Kaplan, R.M., Anderson, J.P., Patterson, T.L., Mccutchan, J.A., Weinrich, J.D., Heaton, R.K., Atkinson, J.H., Thal, L., Chandler, J. and Grant, I. (1995) Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Center. *Psychosom. Medicine* **57**, 138-147.

Kaplan, R.M., Anderson, J.P. and Wingard, D.L. (1991) Gender differences in health-related quality of life. *Health Psycholology.* **10**, 86-93.

Kaplan, R.M., Anderson, J.P., Wu, A.W., Mathews, W.C., Kozin, F., & Orenstein, D. (1989). The Quality of Well-being Scale. Applications in AIDS, cystic fibrosis, and arthritis. *Medical Care, 27,* S27-43.

Kaplan, R.M., Atkins, C.J., & Timms, R. (1984). Validity of a quality of well-being scale as an outcome measure in chronic obstructive pulmonary disease. *Journal Chronic Disease, 37,* 85-95.

Kaplan, R.M., & Bush, J.W. (1982). Health-related quality of life measurement for evaluation research and policy analysis. *Health Psycholology, 1,* 61-80.

Kaplan, R.M., Bush, J.W., & Berry, C.C. (1976). Health status: types of validity and the index of well-being. *Health Services Research, 11,* 478-507.

Kaplan, R.M., Bush, J.W., & Berry, C.C. (1979). Health status index: category rating versus magnitude estimation for measuring levels of well-being. *Medical Care, 17,* 501-525.

Kaplan, R.M., Coons, S.J. and Anderson, J.P. (1992) Quality of life and policy analysis in arthritis. *Arthritis Care Research.* **5**, 173-183.

Kaplan, R.M., Debon, M. and Anderson, B.F. (1991) Effects of number of rating scale points upon utilities in a Quality of Well-Being scale. *Medicine Care* **29**, 1061-1064.

Kaplan, R.M., Ernst, J.A. (1983). Do rating scales produce biased preference weights for a health index? *Medicine Care*, XX1: 193-207.

Katz, J.N., Lasson, M.G., Phillips, C.B., Fossel, A.H., Liang, M.H. (1992) Comparative measurement sensitivity of short and longer health status instruments. *Medical Care*, 30: 917-925.

Katz, J.N., Phillips, C.B., Fossel, A.H., Liang, M.H. (1994). Stability and responsiveness of utility measures. *Medical Care*; 32(2): 183-188.

Keeney, R.L., Raiffa, H. (1976). *Decisions with multiple objectives: preferences and value trade-offs*. John Wiley and Sons, New York.

Kerridge, R.K., Glasziou, P.P., Hillman, K.M. (1995). The use of "quality-adjusted life years" (QALYs) to evaluate treatment in intensive care. *Anaesth Intensive Care*; 23, 322-331.

Kessle, R.C., Foster, C., Webster, P.S., House, J.S. (1992). The relationship between age and depressive symptoms in two national surveys. *Psychology Ageing*, 7: 119-126.

Kind, P. (1990) Measuring valuations for health states: a survey of patients in general practice. Centre for Health Economics Discussion paper 76, University of York.

Kind, P. (1994) An interim tariff for Euroqol health states. Personal Communication.

Kind, P. (1996) The Euroqol instrument: an index of health -related quality of life. In: Spilker, B. (Ed.) *Quality of life and PharmacoEconomics in clinical trials*. 2nd edn. pp. 191-201. Philadelphia, PA: Lippincott-Rivera]

Kind, P., & Dolan, P. (1995). The effect of past and present illness experience on the valuations of health states. *Medical Care, 33,* AS255-AS263.

Kind, P. and Gudex, C.M. (1994) Measuring health-status in the community - a comparison of methods. *Journal of Epidemiology and Community Health* **48**, 86-91.

Kind, P., Gudex, C., Dolan, P. and Williams, A. (1994) Practical and methodological issues in the development of the Euroqol: the York experience. In: Albrecht, G.L. and Fitzpatrick, R. (Eds.) *Advances in Medical Sociology*, pp. 219-253. Greenwich, CT: J A I

Kind, P., Rosser, P., and Williams, A. (1982). Valuation of Quality of life: some psychometric evidence. In Jones-Lee M. W. (ed). The Value of life and safety. Amsterdam: *Elsevier/North Holland.*

Kind, P. and Rosser, R. (1988) The quantification of health. *European Journal Of Social Psychology* **18**, 63-77.

Kind, P., van Dalen, H., Morris, J. and Williams, A. (1993) Comparing Scaling methods: Rosser revisited. Centre for Health Economics Discussion Paper No. 107, University of York.

Krischer, J.P. (1976). Utility structure of a Medical Decision-Making Problem. *Operations Research,* 24: 951-972.

Kroes, E.P., Sheldon, R.J. (1988). Stated Preference Methods: An Introduction. *Journal of Transport Economics and Policy*, 22: 11-25.

Kurtin, P.S., Davies, M.R., Meyer, K.B., et al (1992). Patient-based health status measurements in outpatient dialysis: early experiences inn developing an outcome assessment program. *Medical Care*; 30 (5): MS 136-49.

Lancaster, K. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74: 134-157.

Lancaster, K. (1971). *Consumer demand.* Columbia University Press, New York.

Launois, R., Henry, B., Marty, J.R., Gersberg, M., Lassale, C., Benoist, M. and Goehrs, J.M. (1994) Chemonucleolysis versus surgical diskectomy for sciatica secondary to lumbar disc herniation - a cost and quality-of-life evaluation. *PharmacoEconomics* **6**, 453-463.

Laupacis, A. (1990). The Canadian Erythropoietin Study Group. *British Medical Journal.* 300: 573-578.

Lawrence, K., McWhinnie, D., Goodwin, A., Doll, H., Gordon, A., Gray, A., Britton, J., Collin. (1995) Randomised controlled trial of laporoscopic versus open repair of inguinal hernia: early results. *British Medical Journal,* 311:981-5.

Leighton, J., Read et al (1987). Measuring overall health: An evaluation of three important approaches. *Journal of Chronic Disease*, 40(1): 75-215.

Liang, M.H., Fossel, A.H., & Larson, M.G. (1990). Comparisons of five health status instruments for orthopaedic evaluation. *Medical Care*, *28*, 632-642.

Lichenstein, S., Slovic, P. (1973). Response-Induced Reversals of Preference in Gambling. *Journal of Experimental Psychology*; 101, 16-29.

Lindman, H. (1971). Inconsistent Preferences among Gambles. *Journal of Experimental Psychology*; 89, 390-397.

Lindsay, C.M. (1969). Medical Care and the Economics of Sharing. Economica; 144, 351-362.
- (1973). Real Returns to Medical Education. *Journal of Human Resources*; 8, 331-348.

Lipscomb, J. (1989). Time preference for health in cost-effectiveness analysis. *Medical Care*, *27*, S233-53.

Llewellyn-Thomas, H., Sutherland, H.J., Tibshirani, R., Ciampi, A., Till, J.E., Boyd, N.F. (1982). The measurement of patients' values in medicine. *Medicine Decision Making*, 2: 449-462.
Llewellyn-Thomas, H., Sutherland, H.J., Tibshirani, R., Ciampi, A., Till, J.E., Boyd, N.F. (1984). Describing Health States: Methodological Issues in obtaining values for Health states. *Medical Care*, 22: 543-552.

Lonnqvist, J., Sihvo, S., Syvalahti, E., Sintonen, H., Kiviruusu, O. and Pitkanen, H. (1995) Moclobemide and fluoxetine in the prevention of relapses following acute treatment of depression. *Acta Psychiatr. Scand.* **91**, 189-194.

Lonnqvist, J., Sintonen, H., Syvalahti, E., Appelberg, B., Koskinen, T., Mannikko, T., Mehtonen, O.P., Naarala, M., Sihvo, S., Auvinen, J. and et al (1994) Antidepressant efficacy and quality of life in depression: a double-blind study with moclobemide and fluoxetine. *Acta Psychiatr. Scand.* **89**, 363-369.

Loomes, G. (1993). Disparities between health state measures: is there a rational explanation? In, Gerrard, W. *The Economics of Rationality*. Routledge: London.

Loomes, G. (1995). The myth of the hye. *Journal of Health Economics*, *14*, 1-7.

Loomes, G., Jones-Lee, M.W., Robinson, A. (1994). What do visual analogue scales actually mean? Paper presented to HESG conference, Newcastle.

Loomes, G., & Mckenzie, L. (1989). The use of QALYs in health care decision making. *Social Science and Medicine*, *28*, 299-308.

Loomes, G., Sugden, R. (1982a). Regret Theory: An Alternative Theory of Rational Choice under Uncertainty. *Economic Journal*; 92, 805-824.

Loomes, G., Sugden, R. (1986). Disappointment and Dynamic Consistency in Choice under Uncertainty. *Review of Economic Studies*; 271-282.

Louviere, J. (1988). Conjoint Analysis modelling of stated preferences. *Journal of Transport Economics and Policy*, 22: 93-119.

MacGrimmon, K. (1968). Descriptive and Normative Implications of the Decision Theory Postulates. *Risk and Uncertainty*, Borch, K., Mossin, J. (eds), MacMillan, New York.

Machina, M. (1982). "Expected utility" analysis without the independence axiom. *Econometrica*, 50: 277-323.

Machina, M. (1987). Choice Under Uncertainty: Problems Solved and Unsolved. *Economic Perspectives*; 1, 121-154.

Mackenzie, R., Hollingworth, W. and Dixon, A.K. (1994) Quality of life assessments in the evaluation of magnetic resonance imaging. *Quality of Life Research.* **3**, 29-37.

Magee, T.R., Scott, D.J., Dunkley, A., St Johnston, J., Campbell, W.B., Baird, R.N. and Horrocks, M. (1992) Quality of life following surgery for abdominal aortic aneurysm [see comments]. *British Journal of Surgery.* **79**, 1014-1016.

Majeed, A.W., Troy, G., Nicholl, J.P., Smythe, A., Reed, M.W., Peacock, J. Johnson, A.G. (1996) Randomised, prospective, single-blind comparison of laporoscopic versus small-incision cholesystectomy. *The Lancet*; 347:989-94.

Manzetti, J.D., Hoffman, L.A., Sereika, S.M., Sciurba, F.C. and Griffith, B.P. (1994) Exercise, education, and quality of life in lung transplant candidates. *Journal of Heart Lung Transplantion.* **13**, 297-305.

Margolis, H. (1982). *Selfishness, Altruism and Rationality*. Cambridge, Cambridge University Press.

Marshall, A. (1890). *Principles of economics.* 8th ed. MacMillan, London.

Martin, J., Mettzer, H., Elliot, D., (1988). OPCS Survey of disability in Great Britain Report 1. The Prevalence of disability among adults. London, HMSO.

Matthews, J.N.S., Altman, D.G., Campbell, M.J., Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, 300: 230-5).

Maynard, A. (1991) Developing the health care market. Economic Journal. 101,1277-1286

McCulloch, J., Best, R. (1979). Conjoint measurement: temporal stability and structural reliability. *Journal of Marketing Research*; 16, 26-31.

McCulloch, P and Nelder J.A. (1993) *Generalised Linear Model* Chapman and Hall, London, UK.

McDowell, I., Newell, C. (1987, 1996). *Measuring Health: A Guide to rating scales and questionnaire*. Oxford University Press, Oxford.

McGuire, A., Henderson, J., Mooney, G. (1988). *The Economics of Health Care: An introductory text*. Routeledge and Kegan Paul, London and New York.

McHorney, C.A., Ware, J.E., Lu, J.F.R. et al. (1994). The MOS 36-item Short Form Health Survey (SF-36): III. Tests of data quality, assumptions and reliability across diverse patient groups. *Medical Care*; 32: 40-52.

McHorney, C.A., Ware, J.E., Raczek, A.K.. (1993). The MOS 36-item Short-Form Health Survey (SF-36): II, Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*; 31: 247-63.

McKenna, S., Hunt, S.M., Tennant, A (1993). The development of a patient-completed index of distress from the Nottingham Health Profile: a new measure for use in cost-utility studies. *British Journal of Medical Economics*; 6: 13-24.

Medical Outcomes Trust. (1993). How to score the SF-36 Health Survey. MOS, Boston.

Mehrez, A., & Gafni, A. (1989). Quality-adjusted life years, utility theory, and healthy-years equivalents (published erratum appears in Med Decis Making 1990 Apr- Jun; 10(2):148-9). *Medical Decision Making*, *9*, 142-149.

Mehrez, A., & Gafni, A. (1990). Evaluating health related quality of life: an indifference curve interpretation for the time trade-off technique. *Social Science and Medicine*, *31*, 1281-1283.

Mehrez, A., & Gafni, A. (1991). The healthy-years equivalents: how to measure them using the standard gamble approach. *Medical Deciion Making*, *11*, 140-146.

Mehrez, A., & Gafni, A. (1993). Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress [see comments]. *Medical Decision Making*, *13*, 287-292.

Miller, G.A. (1956). The magical number seven, plus or minus two I some limits on our capacity for processing information. *The Psychological Review*, 63(2): 81-97.

Mitchell, R.C., Carson, R.T. (1989). Using surveys to value Public Goods. *Resources for the future*, Washington D.C.

Miyamoto, J.M., & Eraker, S.A. (1985). Parameter estimates for a QALY utility model. *Medical Decision Making*, *5*, 191-213.

Miyamoto, J.M., & Eraker, S.A. (1988). A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology*, *117*, 3-20.

Miyamoto, J.M., & Eraker, S.A. (1989). Parametric models of the utility of survival duration: tests of axioms in a generic utility framework. *Organ. Behav. Hum. Decis. Proc.*, 44: 162-202.

Mold, J.W., Holtgrave, D.R., Bisonni, R.S., Marley, D.S., Wright, R.A. and Spann, S.J. (1992) The evaluation and treatment of men with asymptomatic prostate nodules in primary care: a decision analysis [see comments]. *J. Fam. Pract.* **34**, 561-568.

Mooney, G.H. *The Valuation of Human Life*. London, Macmillan.

Mooney, G.H. (1986). *Economics, Medicine and Health Care*. Brighton: *Wheatsheaf Books*.

Mooney, G.H. (1994). *Key Issues in Health Economics*. Harvester Wheatsheaf, London.

Mooney, G., Hall, J., Donaldson, C., & Gerard, K. (1991). Utilisation as a measure of equity: weighing heat? *Journal of Health Economics, 10,* 475-480.

Mooney, G., & Lange, M. (1993). Ante-natal screening: what constitutes 'benefit'? *Social Science and Medicine, 37,* 873-878.

Mooney, G., & Olsen, J.A. (1990). QALYs: where next? In A. McGuire, P. Fenn, & K. Mayhew (Eds). *Providing Health Care: The Economics of Alternative Systems of Finance and Delivery*. Oxford University Press, Oxford.

Morris, J., Drummond, M.A. (1989). Category Rating Methods: Numerical and verbal scales-Results from a pilot study. Mimeograph, Centre for Health Economics, University of York, Heslington.

Morrison, G.C. (1994). Consistency within and between methods of health status valuation: a within subject examination of the willingness to pay and Standard Gamble methods. Paper presented at the Econometric Society European Meeting, Maastricht, August.

MVA Consultancy. (1987). *Value of travel time savings*. Institute of Transport Studies, University of Leeds, Leeds.

MVH group. (1994). The measurement and valuation of health: first report on the main survey. Centre for Health Economics, University of York.

MVH group. (1995). The measurement and valuation of health: Final report on the modelling of valuation tariffs. Centre for Health Economics, University of York

Nerenz, D.R., Repasky, Y.P., Whitehouse, M.D. et al (1992). Ongoing assessment of health status in patients with diabetes mellitus using the SF-36 and diabetes TYPE scale. *Medical Care*; 30: 299-319.

Nicholl, J., Brazier, J.E., Milner, P.C. et al (1992). Randomised controlled trial of cost-effectiveness of lithotripsy and open cholecystectomy as treatments for gallbladder stones. *Lancet*; 340: 801-807.

Nord, E. (1989). The significance of contextual factors in valuing health states. *Health Policy, 13,* 189-198.

Nord, E. (1991a). The validity of a visual analogue scale in determining social utility weights for health states. *International Journal of Health Planning and Management, 6,* 234-242.

Nord, E. (1991b). 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 and Medicine, 34,* 559-569.

Nord, E. (1993). Unjustified use of the Quality of Well-Being Scale in priority setting in Oregon. *Health Policy, 24,* 45-53.

Nord, E. (1994). The QALY - a measure of social value rather than individual utility? *Health Economics*, 3(2): 89-93.

Nord, E. (1995). The person-trade-off approach to valuing health care programs. *Medical Decision Making, 15,* 201-208.

Nord, E., Richardson, J., Macarounds-Kichnamm, K. (1993). Social evaluation of health care versus personal evaluation of health states: Evidence on the validity of four health-state instruments using Norwegian and Australian surveys. *International Journal of Technology Assessment in Health Care*, 9(4): 463-78.

Normantaylor, F.H., Palmer, C.R. and Villar, R.N. (1996) Quality-of-life improvement compared after hip and knee replacement. *Journal Of Bone And Joint Surgery-British Volume* **78B**, 74-77.

Nunnally, J.C. (1967). *Psychometric theory.* McGraw-Hill Book Co., New York.

O'Brien, B.J. (1986). *What are my chances doctor? A review of clinical risks.* Office of Health Economics, London.

O'Brien, B.J., Buxton, M.J., & Ferguson, B.A. (1987). Measuring the effectiveness of heart transplant programmes: quality of life data and their relationship to survival analysis. *Journal of Chronic Disease, 40 Suppl 1,* 137S-158S.

O'Brien, B., & Viramontes, J.L. (1994). Willingness to pay: a valid and reliable measure of health state preference? *Medical Decision Making, 14,* 289-297.

O'Hanlon, M., Fox Rushby, J. and Buxton, M.J. (1994) A qualitative and quantitative comparison of the Euroqol and time-trade-off techniques. *International Journal of Health Services.* **5**, 85-97.

Opaluch, J., Swallow, S., Weaver, T., Wessells, C., Wichelns, D. (1993). Evaluating impacts from noxious facilities: Including public preferences in current siting mechanisms. *Journal of Environmental Economics and Management*; 24, 41-59.

Orenstein, D.M. and Kaplan, R.M. (1991) Measuring the quality of well-being in cystic fibrosis and lung transplantation. The importance of the area under the curve. *Chest* **100**, 1016-1018.

Orenstein, D.M., Nixon, P.A., Ross, E.A. and Kaplan, R.M. (1989) The quality of well-being in cystic fibrosis. *Chest* **95**, 344-347.

Orenstein, D.M., Pattishall, E.N., Nixon, P.A., Ross, E.A. and Kaplan, R.M. (1990) Quality of well-being before and after antibiotic treatment of pulmonary exacerbation in patients with cystic fibrosis. *Chest* **98**, 1081-1084.

Organisation for Economic Cooperation and Development (1987) *Financing and Delivering Health Care: a comparative Analysis of OECD countries.* Paris, Social Policy Studies no. 4.

Parducci, A. (1983). Category rating and the relational character of judgement. *Modern Trends in Perception*, Geissler, H.G. (ed). Berlin, VEB Deutcher Verlag Der Wissen-Schaften.

Parkin, D. (1991). Valuing health states: an exploratory data analysis approach. Paper presented to a meeting of the Health Economists Study Group, University of Oxford.

Patrick, D.L., Bush, J.W., & Chen, M.M. (1973a). Methods for measuring levels of well-being for a health status index. *Health Services Research, 8,* 228-245.

Patrick, D.L., Bush, J.W., & Chen, M.M. (1973b). Toward an operational definition of health. *J Health Soc Behav, 14,* 6-23.

Patrick, D.L., Starks, H.E., Cain, K.C., Uhlmann, R.F., & Pearlman, R.A. (1994). Measuring preferences for health states worse than death. *Medical Decision Making, 14,* 9-18.

Pauly, M.V. (1971) *Medical Care at Public Expense*, New York, Praeger.

Pauly, M.V.(1980). *Doctors and Their Workshops*. Chicago, University of Chicago Press.

Pauly, M.V. (1986). Taxation, Health Insurance, and Market Failure in the Medical Economy. *Journal of Economic Literature*; 24, 629-75.

Payne, S.P. and Galland, R.B. (1995) The use of a simple clinical cardiac risk index predictive of long-term outcome after infrarenal aortic reconstruction. *Eur. J. Vasc. Endovasc. Surg.* **9**, 138-142.

Pearmain, D., Swanson, J., Kroes, E., Bradley, M. (1991). Stated preference techniques: a guide to practice. Steer Davis Gleave and Hague Consulting Group, Hauge.

Pereira, J. (1989). What does equity in health mean. Centre for Health Economics Discussion Paper 61, University of York, York.

Pereira, J. (1993). What does equity in health mean. *Journal Of Social Policy, 22*, 19-48.

Petrou, S., Davey, P., & Malek, M. (1992). The application of the Rosser-Kind classification to hip and knee joint replacement surgery.Paper presented to the Health Economists Study Group, University of Sheffield.

Pliskin, J.S., Shepard, D.S., Weinstein, M.C. (1980). Utility functions for life years and health states. *Operations Research*, 28(1): 206-254.

Pratt, J.W. (1964). Risk aversion in the small and in the large. *Econometrica*, 32, 122-136.

Rabin, R., Rosser, R.M. and Butler, C. (1993) Impact of diagnosis on utilities assigned to states of illness. *Journal of the Royal Society of Medicine* **86**, 444-448.

Ramsey, J.B. (1969). Tests for specification Errors in classical linear least squares regression. *Journal of the Royal statistical society*, 31: 350-371.

Rawles, J., Light, J. and Watt, M. (1992) Quality of life in the first 100 days after suspected acute myocardial infarction--a suitable trial endpoint? *Journal of Epidemiology and Community. Health* **46**, 612-616.

Rawls, J. (1971). *A theory of justice*. Harvard University Press. Cambridge.

Read, J.L., Quinn, R.J., Berwick, D.M., & Weinstein, M.C. (1984). Preferences for health outcomes: comparison of assessment methods. *Medical Decision Making, 4*, 315-329.

Read, J.L., Quinn, R.J., & Hoefer, M.A. (1987). Measuring overall health: an evaluation of three important approaches. *Journal of Chronic Disease, 40 Suppl 1*, 7S-26S.

Reed, P.G. (1986) Religiousness among terminally ill and healthy adults. *Res. Nurs. Health* **9**, 35-41.

Revicki, D.A. (1992). Relationship between health utility and psychometric health status measures. *Medical Care, 30*, MS274-82.

Revicki, D.A., & Kaplan, R.M. (1993). Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Quality in Life Research, 2,* 477-487.

Richardson, J. (1994). Cost-utility analysis - what should be measured. *Social Science & Medicine, 39,* 7-21.

Richardson, J., Hall, J., & Salkeld, G. (1990). Cost-utility analysis: the compatibility of measurement techniques and the measurement of utility through time. In C. Selby Smith (ed) *Economics and Health*: 1989. Proceedings of the Eleventh Australian Conference of Health Economists. Public Sector Management Institute, Monash University, Melbourne.

Rissanen, P., Aro, S., Sintonen, H., Slatis, P. and Paavolainen, P. (1996) Quality-of-life and functional ability in hip and knee replacements - a prospective-study. *Quality of Life Research* **5**, 56-64.

Rissanen, P., Aro, S., Slatis, P., Sintonen, H. and Paavolainen, P. (1995) Health and quality of life before and after hip or knee arthroplasty. *Journal of Arthroplasty* **10**, 169-175.

Rosser, R.M. (1988). A health index and output measure. *Quality of life: assessment and application* (ed. S.R. Walker and R.M. Rosser) P133-60, MTP, Lancaster.

Rosser, R., Allison, R., Butler, C., Cottee, M., Rabin, R., & Selai, C. (1993). The Index of Health-related Quality of Life (IHQL): a new tool for audit and cost-per-QALY analysis. In *Quality of life assessment: key issues in the 1990s.*

Rosser, R.M., & Kind, P. (1978). A scale of valuations of states of illness: is there a social consensus? *International Journal of Epidemiolology, 7,* 347-358.

Rosser, R. and Sintonen, H. (1993) The Euroqol quality of life project. In: *Quality of life assessment: key issues in the 1990s*, pp. 197-199.

Rosser, R.M., & Watts, V.C. (1972). The measurement of hospital output. *International Journal of Epidemiolology, 1,* 361-368.

Rutten - Van-Mölken, M. (1994). Costs and effects of pharmacotherapy in asthma and COPD. PhD Thesis, Universitaire Press Maastricht Datawyse Maastricht.

Ryan, M. (1992a) Economic evaluation of In-Vitro Fertilisation: examing the benefits. HERU Discussion Paper No. 13/92, University of Aberdeen.

Ryan, M. (1992b) Stated preferences : a method for establishing the nature of the patient's utility function? HERU Discussion Paper No. 14/92, University of Aberdeen.

Sackett, D.L., & Torrance, G.W. (1978). The utility of different health states as perceived by the general public. *Journal of Chronic Diseases, 31,* 697-704.

Saigal, S., Feeny, D., Furlong, W., Rosenbaum, P., Burrows, E. and Torrance, G. (1994) Comprehensive assessment of the health-related quality of life of extremely low birth weight children and a reference group of children of eight years of age. *Journal of Pediatrics.* **125**, 418-425.

Saigal, S., Rosenbaum, P.L., Furlong, W.J., Feeny, D.H. and Burrows, E. (1995) Self-assessment of their own health-status by extremely low-birth- weight and control teenagers using a multiattribute health-status classification-system. *Pediatric Research* **37**, A 271

Samuelson, P.A. (1947). *Foundations of Economic Analysis*. Cambridge, Mass. Harvard University Press.

Savage, L.J. (1954). *The Foundations of Statistics*. John Wiley, New York, pp. 100-104.

Schoemaker, P.J.H. (1982). The expected utility model: its variants, purposes, evidence and limitations. *Journal of Economic Literature, 20,* 529-563.

Schneiderman, L.J., Kronick, R., Kaplan, R.M., Anderson, J.P. and Langer, R.D. (1992) Effects of offering advance directives on medical treatments and costs [see comments]. *Ann. Intern. Medicine* **117**, 599-606.

Sculpher, M. (1996). Comparing QALYs and HYEs: The case of hysterectomy versus trancervical endometrial rejection. Paper presented to the HESG, January, 1996.

Sculpher, M., Bryan, S., Dwyer, N., Hutton, J. and Stirrat, G.M. (1993) An economic evaluation of transcervical endometrial resection versus abdominal hysterectomy for the treatment of menorrhagia. *British Journal of Obstetrics and Gynaecology* **100**, 244-252.

Selai, C. and Rosser, R. (1995) Eliciting Euroqol descriptive data and utility scale values from inpatients - a feasibility study. *PharmacoEconomics* **8**, 147-158.

Sen, A.K. (1977). Rational fools: a critique of the behavioural founds of economic theory. *Philosophy and Public Affairs*, 6: 317-344.

Sen, A.K. (1977). Social Choice Theory: A Re-examination. Econometrica; 45, 53-90.

Sen, A.K. (1979). Personal Utilities and Public Judgements: or What's Wrong with Welfare Economics'. *Economic Journal*; 89, 537-558.

Sen, A.K. (1980). *Equality of What?* In the Tanner Lectures on Human Values. Cambridge, CambridgeUniversity Press.

Sen, A. (1982). *Choice, Welfare and Measurement*. Oxford, Blackwell.

Sen, A. (1985). *Commodities and Capabilities*. North Holland, Amsterdam.

Sen, A. (1992). *Inequality Re-examined.* Clarendon Press, Oxford.

Shiell, A., Seymour, J., Cameron S. (1995). QALYs, risk-adjusted and HYEs: is there a difference? Paper presented to the HESG conference, University of Aberdeen.

Smith, K., Dobson, M. (1993) Measuring utility values for QALYs: two methodological issues. *Health Economics*; 2:349-355.

Sintonen, H. (1981) An approach to measuring and valuing health states. *Social Science and Medicine* **15C**, 55-65.

Sintonen, H. (1993) [Health-related quality of life measures] Terveyteen liittyvan elamanlaadun mittaamisesta. *Sairaanhoitaja.* 17-19.

Sintonen H (1994a) The 15D measure of HRQoL: reliability, validity, and the sensitivity of it's health state descriptive system. NCFPE Working paper 41, Monash University/The University of Melbourne.

Sintonen H (1994b) The 15D measure of health related quality of life. II Feasibility, reliability, and validity of its valuation system. NCFPE Working paper 42, Monash University/The University of Melbourne.

Sintonen, H., & Pekurinen, M. (1993). A fifteen-dimensional measure of health-related quality of life (15D) and its applications. In Anonymous, *Quality of life assessment: key issues in the 1990s.* (pp. 185-195).

Slovic, P., Tversky, A. (1974). Who Accepts Savage's Axiom? *Behavioural Science*; 19, 368-373.

Sonnenberg, F.A., & Beck, J.R. (1993). Markov models in medical decision making: a practical guide. *Medical Decision Making, 13,* 322-338.

Stevens, S.S. (1966). A metric for the social consensus. *Science*, 151: 530-541.

Stevens, S.S., Galanter, E. (1957). Ratio scales for a dozen perceptual continual. *Journal of Experimental Psychology*. 54: 377.

Stewart, A.L., Hays, R.D., Ware, J.E. (1988). The MOS Short-Form General Health Survey. Reliability and validity in a patient population. *Medical Care*; 26(7): 724-735.

Stewart, A.L., Ware, J. (eds) (1992). *Measuring functioning and well-being*. Duke University Press, Durham and London.

Stewart, M.B., Wallis, K.F. (1982). *Introductory Econometrics*. Basic Blackwell, Oxford.

Streiner, D.L., Norman, G.R. (1989). *Health Measurement Scales: a practical guide to their development and use*. Oxford: Oxford University Press.

Sugden, R. (1980). Altruism, Duty and the Welfare State. In Timms, N. (ed.). *Social Welfare: Why and How*?, London, Routledge and Kegan Paul.

Sugden, R. (1982). On the Economics of Philanthropy. *Economic Journal*; 92, 341-50.

Sugden, R. and Williams, A. (1978). *The Principles of Practical Cost-Benefit Analysis*. Oxford, Oxford University Press.

Sugden, R. (1986). New developments in the theory of choice under uncertainty. *Bulletin of Economic Research*, 38: 1-24.

Swallow, S., Opaluch, J., Weaver, T. (1992). Siting noxious facilities: An approach that integrates technical, economic, and political consideration. *Land Economics*; 68, 283-301.

Tandon, P.K., Stander, H. and Schwarz, R.P., Jr. (1989) Analysis of quality of life data from a randomized, placebo-controlled heart-failure trial. *Journal Clinical Epidemeology* 42, 955-962.

Tarlov, A.R., Ware, J.E., Greenfield, S. et al (1983). The Medical Outcomes Study: An application of methods for monitoring the results of medical care. *Journal of American Medical Association*; 262: 925-30.

Thomas, R., Thomson, K. (1992). Health Related Quality of Life: The Technical Report, University of York, York, UK.

Tolley, G., Kenkel, D., Fabian, R. (1994) *Valuing health for policy: an economic approach.* Chicago, University of Chicago.

Torrance, G.W. (1976). Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Economic Planning Sciences,* 10 (3), 129-136.

Torrance, G.W. (1982). Multi attribute utility theory as a method of measuring social preferences for health states in long-term care. In Kane RL & Kane RA (Eds.), *Values in long-term care.* (pp. 127-156). Levington Books,DC Heath & Co.

Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics*, *5,* 1-30.

Torrance, G.W. (1987). Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases*; *40,* 593-600.

Torrance, G.W., Boyle, M.H., & Horwood, S.P. (1982). Applications of Multi-Attribute Utility Theory to measure social preferences for health states. *Operations Research, 30,* 1043-1069.

Torrance, G.W., Furlong, W., Feeny, D., & Boyle, M. (1995). Multi-attribute preference functions. Health Utilities Index. *Pharmaco Economics*, *7,* 503-520.

Torrance, G.W., Thomas, W.H., Sackett, D.L. (1972). A utility maximisation model for evaluation of health care programs. *Health Services Research,* 7(2): 118-133.

Torrance, G.W., Zhang Y, Feeny, D., Furlong, W., & Barr R. (1992). *Multi-attribute preference functions for a comprehensive health status classification system.* Hamilton, Ontario: Centre for health economics and policy analysis Mc master University.

Torrnace, G.W., Zipursky, A. (1984) Cose0effectiveness analysis of antepartum prevention of Rh immunization. Clinical-perinatol: 11(2): 267-81.

Tramarin, A., Milocchi, F., Tolley, K., Vaglia, A., Marcolini, F., Manfrin, V. and de-Lalla, F. (1992) An economic evaluation of home-care assistance for AIDS patients: a pilot study in a town in northern Italy. *Aids* **6**, 1377-1383.

Tsevat, J., Goldman, L., Lamas, G.A., Pfeffer, M.A., Chapin, C.C., Connors K.F. and Lee, T.H. (1991) Functional status versus utilities in survivors of myocardial infarction. *Medical Care:* **29**, 1153-1159.

Tsevat, J., Solzan, J.G., Kuntz, K.M., Ragland, J., Currier, J.S., Sell, R.L., Weinstein, M.C. (1996). *Medical Care:* 34: 44-57.

Unsworthwhite, J., Kallis, P., Treasure, T. and Pepper, J.R. (1994) Quality-of-life after cardiac-surgery in patients over 70 years of age. *Cardiology In The Elderly* **2**, 133-138.

van-Agt, H.M., Essink-Bot, M.L., Krabbe, P.F. and Bonsel, G.J. (1994) Test-retest reliability of health state valuations collected with the EuroQol questionnaire. *Social Science and Medicine* **39**, 1537-1544.

van Dalen, H., Williams, A. and Gudex, C. (1994) Lay peoples evaluations of health - are there variations between different subgroups. *Journal of Epidemiology and Community Health* **48**, 248-253.

van Hout, B.A. and McDonnell, J. (1992). *Estimating a parametric relation between health description and health valuation using Euroqol Instrument.* In: Euroqol Conference Proceedings, IHE Working Paper 1992:2, Lund, Sweden.

Varian, H.R. (1974). Equity, envy and efficiency. *Journal Economic Theory*, 9: 3-91.

Verhoef, C.G., Verbeek, A.L., Stalpers, L.J. and van Daal, W.A. (1990) [Utility assessment in clinical decision making] Utiliteitsmeting bij de klinische besluitvorming. *Ned. Tijdschr. Geneeskd.* **134**, 2195-2200.

Vikrey, B.G., Mays, R.D., Graber, J. et al (1992). A health-related quality of life instrument for patients evaluated for epilepsy surgery. *Medical Care*; 30: 299-319.

Visser, M.C., Fletcher, A.E., Parr, G., Simpson, A. and Bulpitt, C.J. (1994) A comparison of three quality of life instruments in subjects with angina pectoris: the Sickness Impact Profile, the Nottingham Health Profile, and the Quality of Well Being Scale. *Journal Clinical Epidemeology* **47**, 157-163.

Von Neumann, J., Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton University Press, Princeton.

Wade, D.T. (1991) The q in qalys. *British Medical Journal* **303**, 1136-1137.

Wagstaff, A. (1991a). Qalys and the equity-efficiency trade-off. *Journal of Health Economics, 10,* 21-41.

Wardman, M. (1988). A comparison of Revealed Preference and stated Preference Models of Travel Behaviour. *Journal of Transport Economics & Policy*, 22: 71-91.

Ware, J.E. (1987). Standards for validating health measures: definition and content. *Journal of Chronic Diseases*; 40(6): 473-480.

Ware, J.E., Manning, W.G., Duan, N. (1984). Health Status and the use of outpatient mental health services. *Am. Psychol.* 39: 1090-1100.

Ware, J.E. Sherbourne, C.D. (1992). The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care*: 30: 473-83.

Ware, J.E., Snow, K.K., Kolinski, M., Gandeck, B. (1993). *SF-36 Health Survey manual and interpretation guide*. Boston: The Health Institute, New England Medical Centre, Boston, MA.

Ware, J.E., Jr., Keller, S.D., Gandek, B., Brazier, J.E., & Sullivan, M. (1995). Evaluating translations of health status questionnaires. Methods from the IQOLA project. International Quality of Life Assessment. *International Journal of Technology Assessment in Health Care*, *11*, 525-551.

Watkins, L.D., Bell, B.A., Marsh, H.T. and Uttley, D. (1990) A scale for neurosurgical audit. *British Journal of Neurosurgery.* **4**, 463-465.

Weinstein, M.C. (1988). A QALY is a QALY--or is it? *J Health Econ*, *7*, 289-290.

Weinstein, M.C., Fineberg, H.V. et al (1980). *Clinical Decision Analysis*. Philadelphia: Saunders.

Weisbrod, B.A. (1961). *The Economics of Public Health*. Philadelphia, University of Pennsylvania Press.

Weisbrod, B.A. (1964). Collective-consumption services of individual-consumption goods. *Quarterly Journal of Economics, 78,* 471-477.

Whynes, D.K. and Neilson, A.R. (1993) Convergent validity of two measures of the quality of life. *Health Economics.* **2**, 229-235.

Whynes, D.K., Neilson, A.R., Robinson, M.H. and Hardcastle, J.D. (1994) Colorectal cancer screening and quality of life. *Quality in Life research.* **3**, 191-198.

Wilkin, D., Hallam, L., Doggett, M.A. (1992). *Measures of need and outcome for primary health care*. Oxford: Oxford Medical Press.

Williams, A. (1985). Economics of coronary artery bypass grafting. *British Medical Journal, 291,* 326-329.

Williams, A. (1989). 'Should QALYs be programme specific?' by Donaldson, Atkinson, Bond and Wright. *Journal of Health Economics*, *8,* 485-7; discussion 489-91.

Williams, A. (1992). Measuring functioning and well-being, by Stewart and Ware. Review article, *Health Economics*; 1 (4): 255-258.

Williams, A. (1993). The Euroqol Instrument. A presentation of it's key features at the ESRC/SHHD Workshop on Quality of Life, Edinburgh, unpublished.

Williams, A. (1993). Review of Stewart, A.L., Ware, J.E. (eds.). Measuring functioning and well-being. The Medical Outcome Study Approach, 1992. *Journal of Health Economics*. 1(4): 255-259.

Williams, A. (1995). The measurement and valuation of health: a chronicle.Centre for Health Economics, Discussion Paper, University of York.

Willig, R.D. (1976). Consumer's surplus without apology. *American Economic Review*, 66: 589-97.

Wolfson, A.D., Sinclair, A.J., Bombadier, C., McGreer, A. (1982). Preference measurements for functional states in stroke patients: inter-rater and inter-technique comparisons. In: Kane, R.L., Kane, R.A. (eds.). *Values and Long Term Care*, Lexicon Books; Mass.

Woodhouse, G., Rasbash, J., Goldstain, H., Yang, J., Howarth, J., Plewis, I. (1995) A guide to MLn for New users. University of London, Institute of Education.

World Health Organisation (1948) Constitution of the World health Organisation. Basic documents. Geneva, Switzerland:World Health Organisation.

Wu AW, Mathews WC, Brysk LT, Hampton Atkinson J, Grant I, Abramson I, Kennedy CJ, McCutchan JA, Spector SA and Richman DD (1990) Quality of life in a placebo -controlled trial of Zidovudine in patients with AIDS and AIDS-related complex. *Journal of Acquired Immune Deficiency Syndromes* **3**, 683-690.