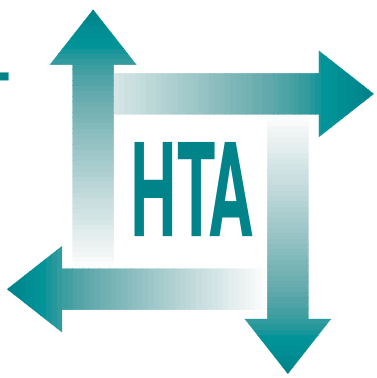
Health Technology Assessment 1999; Vol. 3: No. 9

Review

A review of the use of health status measures in economic evaluation

J Brazier
M Deverill
C Green
R Harper
A Booth

Health Technology Assessment
NHS R&D HTA Programme



Standing Group on Health Technology

Current members

Chair:
Professor Sir Miles Irving,
Professor of Surgery, University of Manchester, Hope Hospital, Salford

Professor Martin Buxton,
Professor of Economics, Brunel University

Professor Francis Creed,
School of Psychiatry & Behavioural Sciences, University of Manchester

Professor Charles Florey,
Department of Epidemiology & Public Health, Ninewells Hospital & Medical School, University of Dundee

Professor John Gabbay,
Director, Wessex Institute for Health Research & Development

Professor Sir John Grimley Evans,
Department of Geriatric Medicine, Radcliffe Infirmary, Oxford

Dr Tony Hope,
The Medical School, University of Oxford

Professor Richard Lilford,
Regional Director, R&D, West Midlands

Dr Jeremy Metters,
Deputy Chief Medical Officer, Department of Health

Professor Maggie Pearson,
Regional Director of R&D, NHS Executive North West

Mr Hugh Ross,
Chief Executive, The United Bristol Healthcare NHS Trust

Professor Trevor Sheldon,
Director, NHS Centre for Reviews & Dissemination, University of York

Professor Mike Smith,
Director, The Research School of Medicine, University of Leeds

Dr John Tripp,
Department of Child Health, Royal Devon & Exeter Healthcare NHS Trust

Professor Tom Walley,
Department of Pharmacological Therapeutics, University of Liverpool

Dr Julie Woodin,
Chief Executive, Nottingham Health Authority

Professor Kent Woods
(**Chair Designate**),
Regional Director of R&D, NHS Executive, Trent

Past members

Dr Sheila Adam,
Department of Health

Professor Angela Coulter,
Director, King's Fund, London

Professor Anthony Culyer,
Deputy Vice-Chancellor, University of York

Dr Peter Doyle,
Executive Director, Zeneca Ltd, ACOST Committee on Medical Research & Health

Professor John Farndon,
Professor of Surgery, University of Bristol

Professor Howard Glennester,
Professor of Social Science & Administration, London School of Economics & Political Science

Mr John H James,
Chief Executive, Kensington, Chelsea & Westminster Health Authority

Professor Michael Maisey,
Professor of Radiological Sciences, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Mrs Gloria Oates,
Chief Executive, Oldham NHS Trust

Dr George Poste,
Chief Science & Technology Officer, SmithKline Beecham

Professor Michael Rawlins,
Wolfson Unit of Clinical Pharmacology, University of Newcastle-upon-Tyne

Professor Martin Roland,
Professor of General Practice, University of Manchester

Professor Ian Russell,
Department of Health Sciences & Clinical Evaluation, University of York

Dr Charles Swan,
Consultant Gastroenterologist, North Staffordshire Royal Infirmary

Details of the membership of the HTA panels, the NCCHTA Advisory Group and the HTA Commissioning Board are given at the end of this report.

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

A review of the use of health status measures in economic evaluation

J Brazier
M Deverill
C Green
R Harper
A Booth

School of Health and Related Research (ScHARR),
University of Sheffield, UK

Published May 1999

This report should be referenced as follows:

Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technol Assess* 1999;**3**(9).

Health Technology Assessment is indexed in Index Medicus/MEDLINE and Excerpta Medica/EMBASE. Copies of the Executive Summaries are available from the NCCHTA web site (see overleaf).

NHS R&D HTA Programme

The overall aim of the NHS R&D Health Technology Assessment (HTA) programme is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and work in the NHS. Research is undertaken in those areas where the evidence will lead to the greatest benefits to patients, either through improved patient outcomes or the most efficient use of NHS resources.

The Standing Group on Health Technology advises on national priorities for health technology assessment. Six advisory panels assist the Standing Group in identifying and prioritising projects. These priorities are then considered by the HTA Commissioning Board supported by the National Coordinating Centre for HTA (NCCHTA).

This report is one of a series covering acute care, diagnostics and imaging, methodology, pharmaceuticals, population screening, and primary and community care. It was identified as a priority by the Methodology Panel and funded as project number 93/47/08.

The views expressed in this publication are those of the authors and not necessarily those of the Standing Group, the Commissioning Board, the Panel members or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for the recommendations for policy contained herein. In particular, policy options in the area of screening will be considered by the National Screening Committee. This Committee, chaired by the Chief Medical Officer, will take into account the views expressed here, further available evidence and other relevant considerations.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

Series Editors: Andrew Stevens, Ruairidh Milne and Ken Stein
Editorial Assistant: Melanie Corris

The editors have tried to ensure the accuracy of this report but cannot accept responsibility for any errors or omissions. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

© Crown copyright 1999

Enquiries relating to copyright should be addressed to the NCCHTA (see address given below).

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.hta.nhsweb.nhs.uk

Contents

List of abbreviations

AQLQ	Asthma Quality of Life Questionnaire
BDI	Beck Depression Inventory*
CBA	cost–benefit analysis
CCA	cost–consequences analysis
CEA	cost-effectiveness analysis
CES	Centre for Epidemiologic Studies
CES-D	Centre for Epidemiological Studies Depression Scale*
CHART	Craig Handicap Assessment and Reporting Technique*
CINAHL	Citation Index for Nursing and Allied Health and Sociofile
CMA	cost-minimisation analysis
COPD	chronic obstructive pulmonary disease
CR	category rating
CUA	cost–utility analysis
DFI	Dyspnea-Fatigue Index
EQoL	expected quality of life*
EUT	expected utility theory
FAI	Frenchay Activities Index*
FEV_1	forced expiratory volume in 1 second
GDS	Global Deterioration Scale*
GHQ	General Health Questionnaire*
HADS	Hospital Anxiety and Depression Scale*
HMQ	Health Measurement Questionnaire
HRQoL	health-related quality of life
HSC	health state classification
HSM	health status measure
HTA	Health Technology Assessment
HUI	Health Utility Index
HYE	healthy year equivalent
IWB	Index of Well-Being*
KPI	Karnofsky Performance Index*
LSI-A	Life Satisfaction Index – A*
LWAQ	Living with Asthma Questionnaire
MAUS	multi-attribute utility scale
MAUT	multi-attribute theory
ME	magnitude estimation
MOS-HIV	Medical Outcome Study – HIV
MOS SF GHS	Medical Outcomes Study Short Form General Health Survey*
MRI	magnetic resonance imaging
MVH	Measurement and Valuation of Health
NEED	NHS Economic Evaluation Database
NHP	Nottingham Health Profile
PTO	person trade-off
QALY	quality-adjusted life-year
QWB	Quality of Well-Being Scale
RAND-36	Medical Outcomes Study Short Form General Health Survey*
RDR	Ruesch Social Disability Rating Scale
RLI	Resident Lifestyle Inventory*
RP	revealed preference
RS	rating scale
RSC	Rotterdam Symptom Checklist*
SAVE	saved young life equivalent
SCI	Science Citation Index*
SF-36	Short-Form 36
SF-6D	Short-Form Six Dimensional
SG	standard gamble
SIP	Sickness Impact Profile
SM	social modifier
SNLAF	Social Network Lifestyle Analysis Form*
SP	stated preference
SSCI	Social Science Citation Index*
TTO	time trade-off
VAS	visual analogue scale
WTP	willingness to pay

* Used only in tables

Executive summary

Background

Health status measures (HSMs)

HSMs are standardised questionnaires used to assess patient health across broad areas including symptoms, physical functioning, work and social activities, and mental well-being. A measure can be disease-specific or generic to any condition, and it can generate a profile of scores, or a single index. The scores can be based on people's preferences (e.g. EQ-5D) or, more usually, arbitrary scoring procedures (e.g. SF-36 assumes equal weighting for most items).

Preference-based HSMs are known as multi-attribute utility scales (MAUSs). These produce a single index score for each state of health which can have a value of 1 or less, where 1 is equivalent to full health and 0 is dead. The scores, known as health state utilities, are used to calculate quality-adjusted life-years. These scores are used in cost–utility analyses.

Scope of the report

This report is concerned with the use of HSMs in economic evaluation, including MAUSs. It does not review all methods of valuing benefits, such as healthy year equivalents, conjoint analysis or willingness to pay.

Objectives

This project reviewed the principles and practice of using HSMs in economic evaluations to develop guidelines for good practice and to identify further research needs.

Methods

Five systematic literature searches were undertaken:

(1) the methodology of using HSMs in economic evaluation
(2) the techniques for valuing health states
(3) the relationship between non-preference-based health measures with preference-based measures
(4) five preference-based measures
(5) the use of HSMs in economic evaluations published in 1995.

Results and conclusions

Judging the appropriateness of HSMs for use in economic evaluation

Conventional psychometric tests of validity were found to be inappropriate, and therefore a checklist was developed to assess the criteria of the practicality, reliability and validity of an HSM which incorporates economists' notion of preferences. The criterion test in economics is agreement with revealed preferences, but such data do not exist in health care. Economic validity can only be examined indirectly using the following:

- the ability to describe health accurately
- the theoretical and empirical bases of the scoring algorithms
- evidence of the measures ability to reflect stated preferences.

A comparison of techniques for valuing health states

The literature relating to the following techniques for valuing health states were reviewed: standard gamble (SG), time trade-off (TTO), visual analogue scale (VAS), magnitude estimation (ME) and person trade-off (PTO). The basic concepts of practicality, reliability, theoretical and empirical validity formed the criteria for reviewing the performance of the valuation techniques.

For practicality and reliability, little evidence relating to ME and PTO techniques was found; with other techniques there is little to choose between them. SG, TTO and the VAS have all proved to be practical on most populations, although VAS techniques have performed slightly better and have cost advantages. There is little difference between the reliability of SG, TTO and the VAS, and present evidence does not offer a basis to differentiate between them. When considering theoretical validity we conclude that only choice-based techniques should be used, that is, SG, TTO and PTO.

Empirical evidence available on the performance of techniques against preferences would suggest that (1) VAS techniques may be measuring aspects of health status rather than valuing health states and (2) choice-based methods are best placed to reflect strength of preference for health states.

Review of preference-based measures of health

The five preference-based measures of health used in economic evaluation – the Quality of Well-Being Scale (QWB), Rosser's disability/distress scale, the Health Utility Index (HUI; mark I to III), the EQ-5D (EuroQoL[c]) and the 15D – were reviewed. The most commonly used measure was the Rosser classification (n = 25), followed by the QWB (n = 24), HUI (n = 10), EQ-5D (n = 8) and 15D (n = 4).

In terms of practicality and reliability, most are brief and easy to use, and four of them can be administered by self-administration. The exception was the QWB, which has a lengthier interview schedule involving detailed probing of the respondents. There was some evidence of the test–retest reliability of the EQ-5D, 15D and HUI-III.

In terms of descriptive validity, the Rosser classification is inferior to the others in its coverage, and has been shown to be less sensitive at detecting health differences than the EQ-5D. The choice from the remaining four depends on the patient group being evaluated and views on the inclusion of social aspects of health. There was evidence of the ability of these measures to detect large differences between patient groups, but they also showed signs of insensitivity to smaller differences.

The QWB, Rosser scale and 15D can be regarded as inferior to the other two measures because their values were not obtained using one of the choice-based techniques. The HUI and EQ-5D use different methods of eliciting weights (SG and TTO, respectively), and there is no consensus amongst health economists as to which is better.

Review of the use of non-preference based measures in economic evaluation

HSMs are not designed for use in economic evaluation, and have a number of problems which make them unsuitable for use in economic evaluations. The main objection is that they do not reflect patient preferences. A poor correlation between HSMs and preference measures was found in published studies. Non-preference-based HSMs can be used to assess the relative efficiency of interventions only in very limited circumstances.

It is recommended that a preference-based measure be used alongside an HSM in trials where it is the intention to undertake an economic evaluation.

Review of economic evaluations conducted in 1995

This review examined the practice of using HSMs in economic evaluations. The number of papers fitting the inclusion criteria for this study (n = 13) suggested that HSMs are not being widely used in economic evaluation.

In most studies, the chosen HSM and the technique of economic evaluation were compatible, and the conclusions presented were legitimate. In many papers, however, there was no information to allow readers of published papers to examine the validity of measures or reasons for choosing it.

Recommendations for research

It is recommended that:

- researchers consider the suitability of their chosen HSM for conducting economic evaluation using the checklist of questions in this report
- the EQ-5D and HUI are currently the best preference-based HSMs, and should be considered for inclusion in all trials intended to be used in economic evaluation
- only choice-based techniques, either SG or TTO, be used to value health states
- SG and TTO values are obtained directly, rather than trying to estimate them from VAS values from a mapping function.

This is a developing field, and the following are priorities for future research:

- a comparison of the EQ-5D and HUI in terms of the features set out in this report
- the estimation of UK preference-based weights for the HUIs and certain key HSMs
- comparisons of MAUSs with other approaches to valuing health benefits
- the development of methods for testing empirical validity of measures for use in economic evaluation
- the empirical validity of the choice-based valuation techniques and their basis in theory.

Chapter 1

Background to the project

Aims and objectives

This report presents the results of a systematic review of the use of health status measures (HSMs) in economic evaluation. This project was commissioned by the NHS Executive's Health Technology Assessment (HTA) methodology panel within the area of: 'assessing different approaches to the measurement of outcomes in HTA and developing recommendations for improvements and standardisation'.

The broad aim of the was to review the principles and practice of using HSMs in economic evaluations, whether or not they were designed for that purpose.

The four stated objectives within the original research proposal were:

- to provide a review of the use of HSMs in economic evaluation
- to compare the performance of preference-based HSMs
- to compare preference measures with HSMs
- to develop guidelines for good practice and identify the needs for further research.

At the request of referees to the original report (submitted in March 1997) the research was extended to include a fifth objective:

- to compare techniques for valuing health states.

The five reviews

Five separate reviews and corresponding search strategies underpin this report. These reviews covered methods and practice, and the general outline of each review is as follows:

- published papers examining the use standardised HSMs in economic evaluation
- published papers on the theoretical arguments for the different techniques of valuing health states and all empirical applications of these techniques to valuing health states
- published studies comparing health status and preference measures
- papers reporting the development and/or use of five preference-based measures of health, known as multi-attribute utility scales (MAUSs)
- published economic evaluations using the frameworks of either cost–utility analysis or CCA published in 1995.

Search methods

Data sources

The core databases used were MEDLINE, EMBASE, the Science Citation Index (BIDS) and the Social Citation Index (BIDS). In addition the general economics databases ECONLIT (SilverPlatter™) and IBIS (British Library Political and Economic Science) were searched. However, the additional yield from the latter two databases was minimal other than confirming the existence of Centre for Health Economics Discussion Papers already known to the team. The NHS Economic Evaluations Database (NEED) was not appropriate for the methodological components of the review but was searched for practical instances of economic evaluation as described below.

Time period covered

MEDLINE was searched back to 1966 (its inception). However, pragmatic cut-off dates varied according to the development of each scale or concept (e.g. EuroQol). The Science Citation Index and the Social Science Citation Index were followed back to 1981 whilst EMBASE covers the period from 1980 onwards. The two general economics databases cover the literature back to the mid-1980s (exact dates vary according to type of material). This gives adequate coverage of the period from which the majority of papers originate. Bibliographic searching was supplemented by citation searching, review of references from the bibliographies of relevant articles and by manual searching of an extensive personal collection of relevant reprints.

Search strategies

The five reviews each have their own search strategy. The search strategies used are fully documented as appropriate, either within the relevant chapters of the report or in an appendix. The search strategies used resulted from a collaboration

between economists (JB, MD and CG) and an expert in searching literature databases (AB).

Structure of the report

The report commences (chapter 2) with an overview of the valuation of health benefits from an economic perspective. It provides a brief description of the techniques of economic evaluation and their principle difference, which is the measure of benefit they use. This is followed by an overview of the theoretical basis for the measures used in economic evaluation, including the quality-adjusted life-year (QALY). The purpose of this chapter is to set the scene for the five reviews presented in the remainder of the report. Chapter 3 is based on the first review and in it we develop a check-list for assessing the relative merits of the various preference-based measures for use in economic evaluations. The aim was to proceed beyond psychometrics and produce a framework which encapsulates the notion of 'validity' from an economist's perspective.

In chapter 4 we present a description and review of the five most widely used methods for valuing health states: the visual analogue scale (VAS; also called the rating scale, RS), magnitude estimation (ME), standard gamble (SG), time trade-off (TTO) and person trade-off (PTO). Chapter 5 appraises five MAUSs: the Quality of Well-Being Scale (QWB), Rosser's disability/distress scale, the Health Utility Index (HUI; mark I to III), the EQ-5D (EuroQoL[c]) and the 15D. This includes a detailed description of the instruments and their uses, and a systematic review against the criteria of practicality, reliability and validity developed in chapter 3 based on the papers identified by the search.

In chapter 6, we undertake a critical review of the use of non-preference-based measures of health-related quality of life (HRQoL) or HSMs in economic evaluation. This includes a theoretical critique and an empirical examination of the relationship between HSMs and preference-based measures using an updated search of studies using both types of measure. On the basis of these recommendations are made regarding the use of HSMs in economic evaluation. We also discuss whether it is possible to develop these measures in ways which make them more suitable for use in economic evaluations.

The results of applying the criteria for judging the validity and suitability of preference- and non-preference-based measures for use in economic evaluation to economic evaluations identified in a literature search of papers published in 1995 are presented in chapter 7.

Finally in chapter 8, we attempt to synthesise the various findings presented in this report to provide guidance on the use of HSMs and to suggest priorities for future research.

Chapter 2

Overview of outcome measurement in economic evaluation

This chapter provides an overview of the subject of outcome measurement in economic evaluation. Its purpose is to set the scene for the remainder of the report. It has been written in a non-technical way designed to be accessible to non-economists.* It is not based on a systematic review of the literature, and we do not claim it to be exhaustive, but it does cover the main issues. The chapter begins by providing a brief description of the techniques of economic evaluation and their associated measures of benefit. This is followed by an examination of the theoretical basis of economic measures of HRQoL used in the economic evaluation of health care.

Techniques of economic evaluation

Economic evaluation is the comparative assessment of the costs and benefits of alternative healthcare interventions (Drummond *et al.*, 1987). The unit for measuring the benefits of health care is the key feature which distinguishes the different techniques of economic evaluation.

Cost-effectiveness analysis (CEA)

This technique compares the cost of alternative ways of achieving a given objective. Where two or more interventions are found to achieve the same level of benefits the least cost intervention is the most cost-effective alternative. This is a cost-minimisation analysis (CMA). Where the benefits of an intervention can be measured by a single dimension, interventions can be compared in terms of their ratio of cost per unit of effect. These effects are usually measured in 'natural' units. The term 'natural' is used to refer to the fact that the measure is unvalued. Typical examples of 'natural' measures used in CEA include life-years saved or number of ulcers prevented. CEAs also use a wide range of surrogate end-points such as detecting cancers, reductions in blood pressure, and improvements in bone mineral density. The important feature of these measures is that more is better than less, on an interval scale over the range being examined. An important question addressed in this review is whether it is possible to conduct a CEA using the scores generated by measures of HRQoL (see chapter 6).

The important characteristic of CEA is that the objective implied by the measure (e.g. detecting cancers) is not being questioned nor its worth valued. In this sense, it is the most straightforward technique of economic evaluation. However, it is also very limited in terms of the questions it can address. It cannot be used to compare interventions which differ in more than one outcome (e.g. where a treatment improves survival at the expense of a poorer quality of life). It is also unable to inform decisions on the efficient allocation of resources between disease groups or healthcare programmes with different outcomes. Nonetheless, it is a widely used technique, which can be extremely helpful in addressing those questions where the objective is not being questioned and no trade-off between outcomes is required.

Cost–utility analysis (CUA)

CUA is like CEA in that it compares interventions in terms of their cost per unit of effect. The difference is that the unit of effect in this case is 'a year in full health', which combines length of life with HRQoL on a single scale. The most widely used measure of 'years in full health' is the QALY.

The number of QALYs is calculated by multiplying a person's life expectancy by the value of the HRQoL experienced in each period as measured by an index score of 1 or less, where 0 is equivalent to death and 1 is full health. Scores can be less than 0 for health states regarded as worse than death. Being on hospital renal dialysis, for example, may be assigned a quality adjustment value of 0.8. A 20 year period on renal dialysis is 16 QALYs, and this is assumed to be equivalent to someone living for 16 years in full health. For more complex health profiles, involving transitions between

* An excellent technical review of the subject can be found in an article by Johannesson *et al.* (1996).

states of health, the QALY score is calculated by summing the product of the time spent in each state and their value.

Healthcare interventions can be compared in terms of their incremental cost per QALY (i.e. the extra cost of an intervention for a given condition group over the next best alternative divided by the extra QALY gain) within and between programmes (Williams, 1985). It even permits comparisons between programmes primarily concerned with increasing survival to those which mainly improve HRQoL. The earliest application of the QALY measure was undertaken in North America by Torrance *et al.* (1977) and in the UK by Williams (1985).

There are two components to the procedure for estimating the quality adjuster for QALYs. The first is a description of the state or profile of a person's health and the second is the valuation of these descriptions. There are different ways of generating the health descriptions, including MAUSs, reviewed in chapter 5. These descriptions are valued using a number of different techniques (as described below).

CUA restricts the benefits of health care purely to gains in health. It is also unable to address the question of how much should be spent on health care compared with other public programmes, with compared to private consumption. It is therefore limited to making comparisons between interventions within the healthcare budget.

Cost–benefit analysis (CBA)

The key feature of CBA is that all the benefits of an intervention are valued in monetary terms. This does not mean that only financial consequences are included, but that non-pecuniary outcomes, such as the effects on survival and HRQoL, have to be valued using money as the numeraire. An intervention is worthwhile if monetary valuation of all the benefits exceeds the costs. This technique can be used to address the question of whether a treatment/programme is worthwhile for society, rather than restricting it to the NHS budget or to a single objective. A further advantage of CBA is that the measure of benefit encompasses a wider range of benefits, and in particular non-health benefits. The theoretical justification for CBA comes from the notion of compensation, which is that those who gain (i.e. the benefits) could compensate the losers (i.e. the costs).*

There are a number of techniques for obtaining monetary valuations of benefits. One is to impute values from people's 'revealed preferences' (RPs) in market settings in order to value benefits. One example of this is in valuing life where the extra earnings of construction workers in risky occupations over safe occupations is used to infer a value for a life. This is regarded as the most appropriate method where it is feasible, since actual decisions are assumed to be a more valid reflection of people's preferences than what someone says they hypothetically would do. However, RP methods are not appropriate in the healthcare field due to the well-documented features of health care, including consumer ignorance and zero or subsidised price at the point of use (Arrow, 1963; Culyer, 1971; Mooney, 1986).

These difficulties have led to the adoption of a range of techniques in applied micro-economics under the broad heading of 'stated preference' (SP) methods or contingent valuation. These methods ask respondents to express how much they would be 'willing to pay' for an intervention, though they are not required to pay.

The use of stated willingness to pay (WTP) has been popular in other areas such as transport and environments, but less so in health economics (Donaldson, 1993). This has arisen in part from a concern about the distributional implications of using WTP, since it assumes the current distribution of income is appropriate. However, there are ways of adjusting for this effect. Another problem arises from the fact that many health systems, such as the NHS, have a fixed budget and hence the decision rule must be modified to examining the relative cost and benefits from different programmes. The use of this technique in health care has recently been revived in the UK (Donaldson, 1995). It is

* 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 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 as compensation, 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.

an important and developing technique, but falls outside the remit of this review.

Cost–consequences analysis (CCA)

In a CCA there is no attempt to combine multiple outcomes into a single indicator of value (such as the QALY). The decision-maker is left with the task of weighing up the costs and the multiple outcomes in a disagreggated form. These outcomes may include a profile of possible outcomes. As such there is no formal theoretical basis for the outcome measures used, and CCA is not strictly one of the techniques of economic evaluation. Despite this, it is a commonly used method of economic evaluation. In the past this method has been known as a 'soft' CBA, but more recently has been called CCA (Drummond, 1994).

The advantage of this approach is that it retains the way of thinking and discipline of economic evaluation. To the extent that the data are helpful, it can be seen within the decision-aiding tradition of economic evaluation (Sugden and Williams, 1978). The disadvantages are that the basis for a decision can often be unclear and will not be based on patient values. The extent to which this analysis can be applied to measures of HRQoL to inform decision-making is examined in chapter 6.

Theoretical basis of economic measures of health

Individual preferences for health

The origin of the economic approach to measuring the benefits of health care can be traced to consumer theory, which is concerned with predicting the choices of individuals between different bundles of commodities (Deaton and Muelbauer, 1980). By commodities, economists mean any potential goods or service which a consumer can purchase. Consumer theory assumes individuals choose the bundle of commodities which maximises their utility subject to their budget constraint, where utility is an indicator of the consumer's strength of preference. Conventional theory postulates that consumers have complete, consistent and transitive preferences over the commodities they consume (i.e. an individual who prefers a bundle of commodities A over B and B over C, will prefer A over C). These restrictions on the nature of people's preferences enable economists to predict how consumers would respond to changes in their income and the prices of different commodities.

An important development in consumer theory has been the recognition that we consume commodities for their characteristics, rather than for their own sake (Lancaster, 1966, 1971). 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 in the future. This investment view of the benefits of health care has been combined with consumer theory (Grossman, 1972).

Applying consumer theory to health, a person deciding whether or not to purchase healthcare services will consider the likely effects they are expected to have on their health and whether the benefits of these effects are worth the costs of the health care. This cognitive process involves some assessment of the value of different aspects of health compared with other goods and services. This may include 'trading', at least implicitly, different aspects of health such as length of life with quality of life (e.g. the decision of whether to have an operation associated with the risk of mortality or life extending chemotherapy with side-effects). Conventional economics uses the amount people are willing to pay in money terms as an indicator of their strength of preference for a good or characteristic of a good, but Buckingham (1993, 1995) and Richardson (1994) have argued that it is also possible to use other numeraires, such as years of life under QALYs.

There is a further complication in health care. Decisions in health care, as in many walks of life, involve uncertainties such as the risks of fatality from common surgical procedures, risks of side-effects from radiotherapy, or the risk of addiction from drug treatments for depression. A technique for predicting individual strength of preference over such uncertain prospects must be based on a theory of decision-making under uncertainty. The key to prediction is being able to make simplifying but reasonable assumptions about human behaviour. The main economic theory of decision-making under uncertainty is expected utility theory (EUT). This theory postulates that individuals choose between prospects (such as different ways of managing a medical condition) in such a way as to maximise their 'expected' utility (Von Neumann and Morgenstern, 1947). Under this theory, for a given prospect such as having a surgical operation, a utility value is estimated for each possible outcome, good or bad. These values are multiplied by their probability of occurring and the result summed to calculate the expected utility of the prospect. This procedure is undertaken for each prospect being considered.

The key assumption made by EUT over and above conventional consumer theory is independence, which means that the value of a given outcome is independent of how it was arrived at or its context. In decision tree analysis this is the equivalent of saying that the value of one branch of the tree is unaffected by the other branches.

Health economists have sought a theoretical foundation for QALYs from EUT (Torrance and Feeny, 1989). For QALYs to accurately reflect preferences, it has been shown that **additional** restrictions must be placed on the nature of individual preferences for health over and above those made for EUT (Pliskin *et al.*, 1980; Miyomoto and Eraker, 1985). QALYs assume that the value of the quality adjuster is constant and unrelated to the duration of a state, when it occurs in time, or where it occurs in relation to other states. This is tremendously important since it means the analyst need only value the states once and can than apply the values to all circumstances. These QALY assumptions can be criticised for being too restrictive (though it must be pointed out that other measures of HRQoL make an even more restrictive assumption about preferences, as will be explained in chapter 6).

There is evidence to suggest, for example, that the value of a health state is altered by the length of time a person spends in the state. Sackett and Torrance (1978) asked patients and members of the general population to value a variety of health states, including hospital dialysis, for durations of 3 months, 8 years and life, and found the mean daily health state utility declined with duration. These results suggest it might be necessary to estimate separate utility values for health states over different durations. Richardson and colleagues (1990) have argued that the utility of a health state may also be 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'.

To overcome the shortcomings of the QALY, Mehrez and Gafni (1991) propose a measure which does not constrain the relationship between quality and quantity and claim that it 'truly' reflects a person's preferences over quantity and quality of life while retaining the intuitive appeal of the concept of the year in full health. They argue it is more consistent with EUT and hence with individuals' preferences. To distinguish it from the QALY they have named their new measure the healthy year equivalent (HYE). It involves the valuation of whole health profiles, which vary in terms of the sequence and duration of health states. This is a more general measure of preferences than the QALY since it makes fewer assumptions. However, it is more complex to estimate values for whole scenarios, and it has been suggested by Johannesson *et al.* (1993) that the HYE (or *ex ante* QALY) 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'. The key question is whether this logistical limitation is outweighed by the advantages from its more general specification of preferences over health. However, there has been remarkably little empirical work on this question.

QALYs imply another restriction on the nature of people's preferences in terms of risk attitude. The outcomes of health care typically involve uncertainty. Even common surgical procedures, for example, are associated with complications, including mortality. A risk-neutral individual would seek to maximise the number of QALYs without any adjustment for risk attitude. Most QALY applications assume risk attitude is neutral. The QALY model has been developed to incorporate a constant attitude to risk. This requires researchers to estimate risk attitude and this can be done by using the SG technique to estimate the quality adjustment and the individual's risk attitude.[*] Another approach to incorporating uncertainty has been to include the probability of different sequences of health events in the scenarios used to estimate the HYEs (e.g. Cook *et al.*, 1994), but this makes the descriptions even more complex.

[*] Non-neutral risk attitudes can be incorporated into the QALY model in the following way (Miyamoto and Eraker, 1985): $U(Q, T) = [V(Q) \times T]^r$. $V(Q)$ is a value function measuring the desirability of state Q and T is the length of time in that state. 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: $U(Q, T) = U(Q) \times T^r$. 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. Miyamoto and Eraker have shown how r can be estimated by ordinary least square analysis from certain equivalent questions (i.e. asking the number of certain years in full health considered equivalent to a gamble involving full health (1) and death (0)).

This review does not seek to resolve the debate about the appropriateness of the assumptions underlying the QALY approach, nor whether HYEs will in practice be an improvement. However, as will be argued in the next chapter, we believe it is important for users of QALY measures to be aware of these assumptions and to consider their likely relevance to the consequences of the intervention they are evaluating.

Social preferences

The conventional view in welfare economics is that social preferences are simply the summation of individual preferences. The compensation test referred to earlier is based on this view. It does not lend support to the notion of QALYs since this measure restricts the individual's utility function to health (as well as making various other assumptions about its specification) (Donaldson, 1995; Mooney, 1994).

However, this 'welfarist' view has been disputed by some economists (Sen, 1985; Culyer, 1989). Culyer (1989), for example, has argued that health care is special and attracts substantial public finance because society has either sympathy (i.e. 'externality') or some moral commitment to the health of others, rather than their utility *per se*. The utility individuals gain from good health is held in higher regard by society than utility from other goods. This would imply a social objective of QALY maximisation, though it would be possible to weight the QALYs on the basis of who receives them in order to reflect social distributional objectives (Wagstaff, 1991). This can be seen as part of the decision-aiding tradition in economics which places less emphasis on conventional welfare economics and utility theory (Sugden and Williams, 1978; Culyer, 1989; Richardson, 1994). It implies that a measure should have a clear meaning to decision-makers so that they feel comfortable using it in choosing between programmes (Richardson, 1994). This 'extra welfare' tradition is not well received by many economists (see the recent review by Johannesson *et al.,* 1996), but it has been important and influential in health economics.

Opinions vary in the health economics literature as to whose values should be elicited. All MAUSs have been valued by samples of the general population, but the valuation of bespoke condition-specific descriptions has often been by patients. 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 Group, 1994). It has been argued that respondents who have experienced the health states are in a better position to understand the states (Buckingham, 1993) and likely to be the most immediate recipients. This would also be consistent with the conventional view in welfare economics, where it is the values of the potential beneficiaries of a given decision which should be used to inform that decision (in order to identify a potential Pareto improvement). Another view is that doctors and other health professionals might be thought to have more experience (though from a third-party viewpoint) of a wider range of health states and hence be in a better position to understand the relative value of different health states. It has also been argued that a representative sample of the general population should be used for informing the allocation of public resources.

Nord (1992) has taken the argument further, and has suggested that given the social objectives of many health systems, the elicitation question should be phrased in terms of the social decision. Under this approach, people are asked to value health programmes from the perspective that the benefits are not to them but to society at large. He claims social values may be different because they incorporate notions of equity and it has even been suggested that the relative weightings of quality of life to survival may be different at a social level. To do this he has developed the PTO technique for valuing health states, and this technique is reviewed in chapter 4.

The practice of measuring preferences for health

There are two components to estimating QALYs. The first involves describing the state or profile of a person's health; the second the valuation of these descriptions.

Eliciting values

Health state valuation techniques require cardinal scale properties to provide information on strength of preference. There are two types of cardinal scales: interval and ratio scales. Both interval and ratio scales provide an ordinal ranking of health states, for example from best to worst, but also provide information on how far apart the health states are. For example, where an intervention is able to move an individual from a health state valued at 0.4 to a health state valued at 0.6, the health gain of 0.2 is said to be the same as an improvement from a health state valued at 0.6 to one valued at 0.8. However, on an interval scale the zero is fixed arbitrarily, and we could not say that the health state valued at 0.8 is twice as good as the health

state valued as 0.4. Interval scales do not allow this type of '*x* times' comparison. Temperature is a common example of values on an interval scale. Ratio scales have zero values which allow comparisons such as 'twice as' or 'half as', and allow us to say that 8 is twice as good as 4. Distance is measured on a ratio scale, thus 8 metres is twice as long as 4 metres. For health state valuations to be used in CUA they require at least an interval scale. For most economic evaluation applications, interval scales are usually regarded as sufficient.

The most commonly used methods to value health states are the VAS (RS), ME, SG, TTO (Torrance, 1986) and PTO (Nord, 1992). These techniques are described in detail in chapter 4. All these techniques have been used to value health states. There are advocates in the economics literature of VAS (Broome, 1993), TTO (Richardson, 1994; Johannesson *et al.,* 1996; Dolan *et al.,* 1996), SG (Feeny and Torrance, 1989; Gafni and Birch, 1993); and PTO (Nord, 1992). The choice of elicitation technique is important because they have been shown to generate different values (e.g. Bombardier *et al.,* 1982; Dolan and Sutton, 1995; Loomes *et al.,* 1995). The relative merits of these valuation techniques are reviewed in chapter 4.

Describing HRQoL for estimating QALYs

Direct utility assessment

One approach has been to use the elicitation techniques described above directly on patients and thereby avoid the need to describe health. This has the logistical advantage of combining two research tasks into one. Perhaps more importantly, people are likely to be better at valuing their own state of health rather than some hypothetical health state. As Buckingham (1993) explains: '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'. The gap between imagination and current experience will partly depend on the accuracy of the health state descriptions. It will also depend on the health experiences of the respondents. A careful selection of respondents who have experienced the health state, or a state like it, would reduce the problem. There might also be a case for approaching those who have witnessed others in such states of health, such as carers or health professionals. However, it excludes the values of other members of society.

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 Erythropoietin 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 (SIP) (which in the past has been criticised for being insensitive (Wilkin *et al.,* 1992)), but the 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 recent study of patients undergoing hip arthroplasty (Katz *et al.,* 1994). Lower responsiveness implies the need for larger sample sizes in order to detect differences and hence 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 exercises that confront patients with some unpalatable scenarios involving, for example death, and hence risk patients withdrawing from a trial. It is usually more acceptable on ethical grounds to collect the descriptive data from patients in a trial, but obtain the values outside of the trial. Furthermore direct utility assessment can only be used for estimating QALYs. The elicitation of HYEs is an *ex ante* valuation of health scenarios, and requires a means of describing health.

MAUSs versus specific descriptions

MAUSs are an important set of instruments for estimating health state values used to calculate QALYs. These are standardised health state classifications (HSCs) with a pre-existing set of preference or utility weights (Drummond *et al.,* 1987). They are widely used in economic evaluations alongside clinical trials to value the benefits of health care. There are a number of MAUSs, and these differ considerably in terms of their dimensions, items and preference weights. As yet, however, there is little guidance in the literature on which to use and to the best of our knowledge, there has been no systematic review of these scales. There are five commonly used MAUSs: the QWB, Rosser's disability/distress classification, the HUI (marks I, II and III), the EQ-5D and the 15D. These are reviewed in chapter 5 in terms of their practicality, reliability and validity.

The other approach is to develop bespoke descriptions of the health states or scenarios experienced by patients receiving different interventions. These are often based on interviews with patients (e.g. Cook *et al.,* 1993), though they could be based on

HSM data. In an early study, Sackett and Torrance (1978) developed descriptions of what it was like to live with chronic renal disease and being treated by one of three regimes: hospital dialysis, home dialysis and renal transplantation. In a cost–utility analysis of breast cancer screening, Hall *et al.* (1992) developed their own description of quality of life with breast cancer because the generic measures were thought to exclude a number of aspects of life found to be important to the women themselves (diagnosis of cancer, physical experience, certain symptoms, etc.).

The debate concerning the appropriateness of specific versus generic descriptions of health is a long-standing one in health services research. In health economics there has been a concern about the relevance and sensitivity of the generic health classification used to derive QALYs (e.g. Donaldson *et al.,* 1988). The appropriateness of a generic health classification depends on the condition and for some conditions, studies have found generic measures to be as sensitive as condition-specific measures (Fitzpatrick *et al.,* 1993). On the other hand, generic classifications are usually easier to use (Gerrard, 1992). A generic classification has a set of off-the-shelf values, whereas condition-specific descriptions will have to be re-constructed from trial data and then valued as part of the study (Brazier and Dixon, 1995). The use of a generic measure also improves comparability between studies, and hence it could be argued, is more suitable for making cross-programme comparisons. In theory, the results of studies using condition-specific descriptions should be comparable since they are using the common numeraire of the QALY. However, the respondents used to generate the values will be different, and hence less comparable. A generic classification also has the advantage of being able to define what aspects of quality of life are important for informing the allocation of public funds.

Conclusions

An economic evaluation is the comparative assessment of the costs and benefits of healthcare interventions. The purpose is to generate information that will assist decision-makers to determine the most efficient way of allocating their scarce resources between competing demands. Economic evaluation raises a host of theoretical and methodological problems for researchers in the design of, data collection for and analysis of studies. This report is concerned with one set of problems, namely the assessment of benefits, and focuses on the use of measures of HRQoL in economic evaluation.

The purpose of this chapter was to set the report in context. It describes the different approaches, including those which are excluded from this review which deserve attention in their own right: HYEs, WTP and conjoint analysis. This chapter has also provided the theoretical background to the economic methods reviewed in this report and important reference material for the remainder of this report.

References

Arrow KJ. Uncertainty and the welfare economics of medical care. *Am Econ Rev* 1963;**53**:941–73.

Brazier JE, Dixon S. The use of condition specific outcome measures in economic appraisal. *Health Econ* 1995;**4**:255–64.

Broome J. QALYs. *J Public Econ* 1993;**50**:149–67.

Buckingham K. A note on HYE (healthy years equivalent). *J Health Econ* 1993;**12**:301–9.

Buckingham K. Economics, health and health economics – HYEs versus QALYs – a response. *J Health Econ* 1995;**14**:397–8.

Cook J, Richardson J, Street A. A cost–utility analysis of treatment options for gallstone disease – methodological issues and results. *Health Econ* 1994;**3**:157–68.

Culyer AJ. The nature of the commodity health care and its efficient allocation. *Oxford Economic Papers* 1971;**24**:189–211.

Culyer AJ. Commodities, characteristics of commodities, characteristics of people, utilities and quality of life. In: Baldwin S, Godfrey C, Propper C, editors. The quality of life: perspectives and policies. London: Routledge, 1989.

Deaton A, Muellbauer J. Economics and consumer behaviour. Cambridge: Cambridge University Press, 1980.

Donaldson C. Theory and practice of willingness to pay for health care. HERU Discussion Paper 01/93. Aberdeen: University of Aberdeen, 1993.

Donaldson C, Atkinson A, Bond J, Wright K. Should QALYs be programme-specific? *J Health Econ* 1988;**7**:239–57.

Donaldson C, Hundley V, Mapp T. Willingness to pay: a new method for measuring patients' preferences? HERU Discussion Paper, University of Aberdeen, 1995.

Drummond M. Economic analysis alongside controlled trials: an introduction for clinical researchers. Leeds: DoH, 1994.

Drummond MF, Davies L. Economic analysis alongside clinical trials: revisiting the methodological issues. *Internat J Tech Assessment in Health Care* 1991;**7**(4):561–73.

Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 1987.

Feeny DH, Torrance GW. Incorporating utility based quality of life assessment measures in clinical trials. *Medical Care* 1989;**27**(3): S190–204.

Fitzpatrick R, Zeibland S, Jenkinson C. *et al.* A generic health status instrument in the assessment of rheumatoid arthritis. *Br J Rheum* 1992;**17**:439–47.

Gafni A, Birch S. 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* 1993;**23**:219–28.

Gafni A, Birch S, Mehrez A. Economics, health and health economics – HYEs versus QALYs. *J Health Econ* 1993;**12**:325–39.

Grossman M. On the concept of human capital and the demand for health care. *J Political Economy* 1972;**80**:223–55.

Gerard K. Cost–utility in practice – a policy makers guide to the state-of-the-art. *Health Policy* 1992;**21**:249–79.

Hall J, Gerard K, Salkeld G, Richardson J. A cost utility analysis of mammography screening in Australia. *Soc Sci Med* 1992;**34**:993–1004.

Hicks JR. The foundations of welfare economics. *Economic J* 1939;**49**:696–710.

Johannesson M, Jonsson B, Karlsson G. Outcome measurement in economic evaluation. *Health Econ* 1996;**5**:279–96.

Kaldor N. Welfare propositions of economics and interpersonal comparisons of utility. *Econ J* 1939;**49**:549–52.

Katz JN, Phillips CB, Fossel AH, Liang MH. Stability and responsiveness of utility measures. *Med Care* 1994;**32**(2):183–8.

Kind P, Rosser P, Williams A. Valuation of quality of life: some psychometric evidence. In: Jones-Lee MW, editor. The value of life and safety. Amsterdam: Elsevier/North Holland, 1982.

Lancaster K. A new approach to consumer theory. *J Pol Econ* 1966;**74**:134–57.

Laupacis A. The Canadian Erythropoietin Study Group. *BMJ* 1990;**300**:573–8.

Loomes G, Mckenzie L. The use of QALYs in health care decision making. *Soc Sci Med* 1989;**28**:299–308.

Mehrez A, Gafni A. The healthy-years equivalents: how to measure them using the standard gamble approach. *Med Decis Making* 1991;**11**:140–6.

Miyamoto JM, Eraker SA. Parameter estimates for a QALY utility model. *Med Decis Making* 1985;**5**:191–213.

Mooney GH. Economics, medicine and health care. Brighton: Wheatsheaf Books, 1986.

Mooney GH. Key issues in health economics. London: Harvester Wheatsheaf, 1994.

Nord E. Methods for quality adjustment of life years. *Social Sci Med* 1992;**34**:559–69.

Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res* 1973;**8**:228–45.

Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health states. *Operat Res* 1980; **28**(1):206–54.

Richardson J. Cost–utility analysis – what should be measured. *Soc Sci Med* 1994;**39**:7–21.

Richardson J, Hall J, Salkeld G. Cost–utility analysis: the compatibility of measurement techniques and the measurement of utility through time. In: Selby Smith C, editor. Economics and health: 1989. Proceedings of the 11th Australian Conference of Health Economists. Melbourne: Public Sector Management Institute, Monash University, 1990.

Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chronic Dis* 1978;**31**:697–704.

Sen A. Commodities and capabilities. Amsterdam: North Holland, 1985.

Sugden R, Williams A. The principles of practical cost–benefit analysis. Oxford: Oxford University Press, 1978.

Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Econ Planning Sci* 1976;**10**(3):129–36.

Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986;**5**:1–30.

Von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton: Princeton University Press, 1944.

Wagstaff A. QALYs and the equity–efficiency trade-off. *J Health Econ* 1991;**10**:21–41.

Weistein MC, Statson WB. Foundations of cost-effectiveness analysis for health and medical practice. *New Engl J Med* 1977;**296**:716–21.

Williams A. Economics of coronary artery bypass grafting. *BMJ* 1985;**291**:326–9.

Wilkin D, Hallam L, Doggett MA. Measures of need and outcome for primary health care. Oxford: Oxford Medical Press, 1992.

Chapter 3

A check-list for judging preference-based measures of health for use in economic evaluation

There have been a number of published reviews of measures of HRQoL (e.g. see Streiner and Norman, 1989; McDowell and Newell, 1989; Wilkin *et al.,* 1992; Bowling, 1992). The absence of economic considerations from the criteria used in these reviews has often resulted in economic measures of HRQoL being neglected and portrayed as 'invalid' or irrelevant in the assessment of health benefits. This is an important omission given the role of assessing efficiency in modern health services research.

Our aim in this chapter is to fill this gap in the current work by developing a check-list for judging the merits of preference-based measures (such as the QWB or the EQ-5D) by adapting the criteria used by psychometricians to judge the performance of non-preference-based measures of health status (HSMs) such as the Short-Form 36 (SF-36) health survey or the Nottingham Health Profile (NHP). This check-list should be useful to researchers reviewing different measures of health benefit, whether they be generic multi-attribute scales or more condition-specific scenarios, for use in economic evaluation.

We begin by reviewing the conventional psychometric criteria used by health services researchers for assessing measures of health in order to examine what lessons can be learnt from this tradition and to highlight where they diverge from the economic requirements of a measure. This is followed by a section on developing an economic understanding of validity. We then build upon the foundations laid by the psychometric tradition by developing a check-list of questions to ask of any measurement instrument being considered for use, or being used, in an economic evaluation. The concluding section considers the uses of the check-list.

Search strategy and methods of review

A systematic search of the data sources described in chapter 1 was undertaken using the search strategy shown in *Box 1*.

BOX 1 Search strategy
Health status measure*
Health status questionnaire*
Health status indicator*
Quality of life

The above are very broad concepts that result in good sensitivity but poor specificity. In the case of 'quality-of-life' it was recognised that there is a significant body of literature in disciplines other than health economics. It was therefore decided to improve the specificity of retrieval of this concept by combining this term (as both a free-text phrase and a designated index term) with the terms presented in *Box 2*. This strategy was then translated into appropriate search terms for subsequent searches on other databases.

BOX 2 Search strategy (using MEDLINE as an exemplar)
\economics as a subheading
costs-and-cost-analysis {exploded} as a medical subject heading
economic* as a text word
cost* in title, abstract or subject headings

This strategy found over 1300 potential papers for review. The abstracts of these papers were screened to ascertain their relevance. The process identified 154 for review, and these have been listed at the end of this chapter.

These papers have not been systematically reviewed in the conventional sense of applying an established methodology as used by the Cochrane groups. It is not possible to grade against quality criteria, since such a method of grading does not exist other than the peer review system used by journal editors. Nor is there a quantitative method for assessing opinion. This is intended to be a comprehensive review, and one which presents an accurate balance of opinion (we have tried to

reflect disagreements rather than hide them) from the economics literature, but it inevitably contains our own judgements and opinions.

Psychometric criteria – what can economists learn?

The psychometric approach of measuring health was originally derived from a field of enquiry known as psychophysics, which attempted the measurement of human perceptions of different stimuli such as heat and light (Nunnally, 1967). Psychometrics extended psychophysics to more subjective concepts such as intelligence, attitudes and health perception. The methods of psychometrics have been applied widely in health measurement (McDowell and Newell, 1987) and were integral to the construction and testing of health status questionnaires such as the SF-36 (Ware and Sherbourne, 1992; Stewart and Ware, 1992; McHorney *et al.,* 1994).

The psychometric literature provides a set of criteria for assessing the performance of an instrument. The most commonly used are practicality, internal consistency, reliability, validity and responsiveness to change in health. Researchers have developed a variety of empirical methods for testing the performance of an instrument against these criteria (Streiner and Norman, 1989). We examine each of these criteria in turn in terms of their relevance in assessing measures for use in economic evaluation.

Practicality

An instrument must be acceptable to the patient and to those representing the interests of the patient, such as healthcare professionals and ethics committees. The length of time it takes to administer an instrument has implications for feasibility, cost and where there is respondent fatigue, the quality of data. This is an important consideration for any measure. Some economic instruments present additional problems. The impact of the length, difficulty and acceptability of an instrument can be assessed quantitatively in terms of the proportion of those approached who agree to participate (i.e. the response rate) and the level of missing data (i.e. completion rate).

Internal consistency

For measures of HRQoL, internal consistency (or internal reliability as it is sometimes known) has tended to be assessed in psychometrics in terms of the homogeneity of items within a dimension. Consistency is often used in health economics to refer to the extent to which respondents' valuations correspond with the known logical ordering of health states or the underlying assumptions about preferences over health. This notion of consistency is addressed later in this chapter. The items or questions in an instrument are assumed to tap a particular dimension of health and therefore responses to items in the same dimensions should be correlated with one another. This definition of internal consistency could conflict with the requirements of a measure for economic evaluation since it may result in the exclusion of items which do not fit neatly into one of the hypothesised dimensions but are important in terms of patient or societal preferences. This would also be a concern in psychometrics in terms of content validity, as we discuss below.

Reliability

A measure must be able to reproduce a series of results over repeated measurements on an unchanged population with the minimum amount of random error. Reliability includes stability over time (retest reliability), agreement between raters (inter-rater reliability), and agreement between scores obtained from different places of administration. All measures have some degree of random error, and the consequence of greater random error is the need for larger sample sizes. Reliability is important for any measurement instrument, including economic measures of outcome such as QALYs (Torrance, 1986; Froberg and Kane, 1989; Dolan *et al.,* 1996).

Validity

Validity has been defined as the extent to which an instrument measures what it is intended to measure. The validation process is therefore concerned with seeking to establish the extent to which a measure serves the purpose for which it is being used. Ideally an instrument would be tested against a criterion or 'gold standard'. In the absence of such a gold standard measure for HRQoL, psychometricians have developed various indirect ways of establishing validity. The most commonly used are content validity, face validity, construct validity and concurrent or convergent validity (Streiner and Norman, 1989; Wilkin *et al.,* 1992). Each of these is examined below.

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). No measure can cover all dimensions and include every conceivable item, and there is inevitably a trade-off between completeness and

parsimony. Claims for content validity typically rest on the comprehensiveness of the instrument and the methods used to generate its dimensions and items. The need for comprehensiveness can act as a constraint on the application of internal consistency as described above.

Developers of the first version of the EuroQol instrument used the content of existing health status questionnaires (HSMs) (EuroQol Group, 1990). This 'expert' approach to generating dimensions and items could be criticised for not accurately reflecting the views of those concerned, such as the patients, their carers or the health professionals. In health services research, there is an interest in using the views of patients in the development of the instruments measuring health. The approach used by some methodologists, such as the developers of the NHP, was to obtain an initial pool of statements from interviews with patients (Hunt *et al.,* 1986). Economists, being concerned with ensuring the measure correctly reflects the arguments of an individual's utility function, are likely to prefer this patient-based approach.

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 regarded as inappropriate. This is important for the acceptance of a questionnaire and whether it is likely to generate valid descriptive data. This is a subjective test to be undertaken by the researcher, and may include consulting relevant health professionals, or the patients themselves. This is also going to be important for use in an economic evaluation.

Construct validation

Construct validation represents a series of procedures for testing the validity of an instrument. The procedures are all concerned with assessing the extent to which the instrument correlates with other hypothesised measures or indicators of the health concept or concepts of interest. There are two commonly used approaches:

- **Group comparisons**. This is where a measure is judged in terms of its ability to differentiate between groups thought to differ in terms of their health. However, the 'constructs' conventionally used to test measures of HRQoL may not reflect preferences. Age, for example, is associated with health, but it cannot be assumed that older people would give a lower valuation for their own health state. Clinical opinion on the severity of a condition may be poorly correlated with patients' views. Tests of construct validity must be made appropriate for the measurement of preferences.
- **Convergent validity**. This is the extent to which a measure correlates with another measure of the same concept. The comparator instrument would usually be an existing and widely accepted measure of health. Again this would not be appropriate without adaptation for preference-based measures.

Responsiveness

The concept of responsiveness is closely related to validity. Responsiveness is the ability of an instrument to measure clinically significant changes in health (Wilkin *et al.,* 1992). It is regarded as the key property of a measure for evaluating the impact of healthcare interventions. It is also related to reliability since the more stable a measure, the more able it is to detect change. Responsiveness is usually assessed statistically using measures such as the 'effect size', where the mean change in score is divided by either the standard deviation at the baseline or the standard deviation of the change (Guyatt, 1985). The effect size indicates the relative size of the 'signal' in comparison to underlying 'noise' in the data. The effect sizes of different instruments are compared for groups of patients assumed to have experienced a health change, such as after an operation of known effectiveness (e.g. a knee operation) or where the patient's doctor reported a change in health.

A common assumption in the assessment of responsiveness is that for a given health change, the HSM with the larger effect size is the better measure (Guyatt, 1985; Fitzpatrick *et al.,* 1992; Katz *et al.,* 1994). Where the objective is to minimise sample size this makes sense. However, when the purpose is to compare the size of change between treatments as part of an economic evaluation, within or between conditions, it is the value of change which matters. Effect sizes do not indicate value or the importance of a change. It is the sensitivity of or responsiveness of an instrument which is important for economic evaluation.

Overview

There are important lessons for economics from the psychometrics literature. Practicality and reliability are important criteria for assessing the performance of any instrument and are concerns which should be common to economists and psychometricians alike. The importance of practicality is compounded in economic evaluations conducted alongside clinical

trials where the instrument will be an additional burden (usually for the patient) to other measures of health. Reliability has major implications for sample size calculations and hence the cost of conducting a trial (O'Brien and Drummond, 1993). Amid the theoretical debates currently taking place in health economics journals it is important not to lose sight of the importance of practicality and reliability in assessing the value of an instrument for use in economic evaluation (Dolan *et al.,* 1996; Froberg and Kane, 1989; Torrance, 1986). These criteria are however irrelevant unless the instrument is measuring the right concept, namely preferences.

The content and face validity of the items of a questionnaire also appear to be promising candidates for a check-list. and should not be ignored. However, we found the psychometric criteria of internal consistency, construct validity, and responsiveness on their own to be inappropriate for judging the suitability of a measure for use in economic evaluation. This arises from a fundamental difference in what psychometricians and economists are seeking to measure. Whilst psychometricians are seeking to measure or numerically describe patient perception along different dimensions of health for clinical trials and routine monitoring, economists want to know the **relative value** patients and others place on the dimensions and their components in order to undertake more than the most rudimentary form of economic evaluation. The value of a health improvement will be related to a measure of the size of the change, but these two concepts will not be perfectly correlated. For example, someone may regard a large health improvement (such as the ability to walk upstairs) as being of little or no benefit if they live in a bungalow. Conversely an apparently small improvement in pain may be highly valued by the patient.

Validity is judged in terms of the extent to which an instrument measures what it is intended to measure and economists are interested in measuring the value placed on health rather than in measuring health *per se*. The rest of this chapter therefore considers how economists would want to define and therefore test the validity of measures of health used in economic evaluations.

Towards an economic understanding of validity

Economic evaluation requires a measure of benefit which reflects individual or societal preferences. Even the least sophisticated technique of economic evaluation of CMA requires a measure to be related to preferences, though it need only have ordinal properties and hence be able to rank states of health from best to worst in the right order. For the more sophisticated techniques of CUA and CBA, the measure must reflect preferences on a cardinal scale. The intervals of the scale must be equal, and hence we can say, for example, that a movement from say 4 to 3 is equal to a movement of 2 to 1. This enables healthcare programmes to be compared in terms of a ratio of cost per unit of change.

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. However, RP methods have not been applied in the healthcare field due to the well-documented features of this commodity (Arrow, 1963; Culyer, 1971; Donaldson and Gerard, 1993). RP methods require the consumer to be sovereign, but in health care the consumer is often ignorant of the outcomes of care. Furthermore, the doctor can act as the patient's agent in the consumption of health care, but the level of ignorance is such that the patient cannot be sure his/her doctor is being a perfect agent. It cannot be assumed that the health services provided would have been the consumer's preferred choice.

There are a variety of views in the health economics literature on testing measures of validity. One approach is to be sceptical about the value of trying to prove validity at all. This view is reflected in a comment by Williams (1995), who suggested that '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 other 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 the preference-based measures was reviewed in chapter 2, where it was shown that there has been a considerable amount of disagreement on which is the most theoretically correct measure for use in economic evaluation. Indeed, it is the debates about theory which have dominated the economics literature on outcome measurement.

There is a further question as to whether social resource allocation decisions should be informed by some aggregation of individual preferences (Loomes and McKenzie, 1989). Testing the validity of social values would imply a different approach. Nord (1993) suggested an approach whereby: '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 has used this question with members of the general population. Within the broader decision-aiding tradition, the measure should have a clear meaning to decision-makers so that they feel comfortable using it in choosing between programmes (Richardson, 1994), but no formal method for testing validity has been proposed.

Such a diversity of opinion makes it difficult to suggest a single set of criteria for assessing the validity of a measure for use in economics evaluation. Nonetheless, we believe it is important to consider explicitly the validity of preference-based measures of health in designing, conducting and reviewing an economic evaluation. We have therefore developed a check-list for assessing their validity which includes the agreements and disagreements within the economics literature. The list examines two parts of a measure: firstly the description of the consequences and secondly their valuation. A critical assessment of these two parts should help in understanding the extent to which an instrument is **able** to be a valid cardinal measure of preferences validity. We would also not wish to loose sight of empirical validity, and this forms the third part of the check-list. These three parts of the list are now discussed in turn.

Descriptive validity

To be confident that the values generated by a measure reflect preferences, they must be generated from accurate descriptions of health or changes in health of relevance to a person's utility function. In contrast to psychometrics, there has been little written in the economics literature about this very important aspect of outcomes measurement. Published economic evaluations rarely address the issue and yet it has been suggested that quite small differences in the content of health state descriptions can alter the results substantially (Smith and Dobson, 1993). There has also been understandable concern about the relevance and sensitivity of the health classifications used to derive QALYs (Donaldson *et al.,* 1988; Hall *et al.,* 1992; Carr-Hill and Morris, 1991). The descriptive systems of the Rosser disability and distress classification, for example, have been found to be insensitive compared to other measures of health (Hollingworth *et al.,* 1995)

It has been suggested that condition-specific descriptions should be used instead of generic ones since they can be made more relevant to the condition and hence more sensitive (Donaldson *et al.,* 1988). However, this is not always the case. For example, in a study of patients with rheumatism the generic EQ-5D was found to be as sensitive to health differences and changes as instruments designed for this patient group. The use of specially constructed disease-specific scenarios in the derivation of QALYs or HYEs is no guarantee of descriptive validity. In a study published by Cook *et al.* (1994) the vignettes were very simplistic and not able to describe the diversity of outcomes found in prospective study of patients receiving the same treatments. For example, the following description was used to describe a successful laporoscopic procedure: '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'.

One way of avoiding the need to describe health states or scenarios is to obtain preference data directly from patients experiencing the healthcare intervention. However, as argued in chapter 2, there are reasons why this may not be desirable and besides, it is often not possible to use preference elicitation techniques directly on patients. The assessment of descriptive validity is therefore an essential step. The question is how it should be done. We suggest that content and face validity, together with construct validation, be used to test the descriptions of the classifications.

Content and face validity

The psychometric criteria of content and face validity, though subjective, are nonetheless important to assess the comprehensiveness, relevance and sensitivity of the dimensions in MAUSs or scenarios. The content of a measure in terms of its dimensions and items implicitly defines the contents of a utility function. Economists who are concerned with ensuring the measure correctly reflects the arguments of an individual's utility function, that is, the things which individuals value, may prefer a direct method of eliciting patients views when generating items and dimensions.

Construct validity

It is important to have an empirically based means of testing the descriptive validity of an instrument.

Construct validation is appropriate for testing the validity of the description of health or health change underlying a MAUS. The ability of an instrument to reflect known or expected differences and changes in health is an essential precursor to its ability to reflect preferences. Otherwise there is a danger that the failure of a score to detect a difference is incorrectly interpreted to imply that the descriptive component of the instrument is insensitive. The score of a preference-based measure may fail to detect the difference simply because the difference is not valued by patients. On the other hand, it could be the result of an insensitive scoring system, as was shown to be the case with the Rosser classification

The construct validation of a health scenario would involve different tests. Developers would have to demonstrate the empirical accuracy of the scenarios against evidence from trials. Potential users would have to show how the scenarios applied to their own patient/treatment group.

The methods of valuation

There are four aspects of the methods of valuation to be addressed: the question of whose values to elicit, the assumed model of preferences underlying the method of valuation, the technique of valuation, and the quality of the valuation data.

Whose values?

Views in the literature vary as to whose values should be incorporated into an evaluation. This is an important judgement since there is evidence of valuations varying by disease experience, age and education (e.g. Sackett and Torrance, 1978; Slevin *et al.,* 1990; MVH Group, 1994). It has been argued that respondents who have experienced the health states are in a better position to understand the states (Buckingham, 1993) and likely to be the most immediate recipients. This would also be consistent with the conventional view in welfare economics, where it is the values of the potential beneficiaries of a given decision which should be used to inform that decision (in order to identify a potential pareto improvement). Another view is that doctors and other health professionals might be thought to have more experience (though from a third party viewpoint) of a wider range of health states and hence be in a better position to understand the relative value of different health states. It has also been argued that a representative sample of the general population should be used 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 question of whose preferences or values should be used in valuation surveys is ultimately a question of whose perspective is regarded as relevant for a given decision. These is currently no concensus in the health economics literature as whose values should be used to inform resource allocation decisions.

Assumptions about preferences

QALYs and HYEs rely on making assumptions about the nature of people's preferences for health (Johannesson *et al.,* 1996). The QALY 'model' makes the most restrictive assumptions. It assumes that the value of a health state is independent of when it occurs, its duration, and its context (chapter 2). In most applications of the QALY, decision-makers are also assumed to be risk neutral. Gafni *et al.* (1993) argue that users of QALYs should **prove** the additional assumptions made by the QALY. This argument could be applied to the more general assumptions underlying QALYs and HYEs of EUT.

Unfortunately, there is little evidence on the importance of any violations of the assumptions of QALYs and HYEs in different circumstances. Nonetheless, it should be incumbent upon the potential user to consider whether these assumptions are likely to be appropriate for the intervention/condition they are planning to evaluate. In the case of the QALY, the researcher should be able to gauge the potential for significant departures from the assumptions of the QALY and their likely importance; for example, whether the patient is likely to adapt to the health state through time or whether prognosis might influence the value of a health state (Richardson *et al.,* 1990). For applications of the QALY model, the importance of risk in the outcomes of an intervention should be considered. A small risk of mortality was found by Cook *et al.* (1994) to be weighted more heavily than is predicted by the QALY model. This would suggest the risk-neutral QALY model is inappropriate in such circumstances. The more general HYE model proposed by Mehrez and Gafni (1989) is less restrictive than the QALY model, but it nonetheless assumes risk neutrality with respect to healthy years and the axioms of EUT.

Requiring the researcher to consider the appropriateness of the assumptions of their chosen method of valuation might be thought to be overly demanding. Some of the issues are quite complex. However, we believe it is important for empirical work to start to reflect the concerns being raised in the literature about some of the methods. It does not seem unreasonable to expect someone who is designing and conducting an economic evaluation

to at least consider the likelihood of any violations and how important they could be in altering the conclusions of the study. Users of the results of such a study need not undertake such an exercise, but they should be given appropriate guidance by the researchers in the interpretation of the results.

It should be noted that whilst these theoretical issues preoccupy the journals of health economics, they seem to have largely been ignored in the psychometric literature. HSMs do not usually incorporate preferences in an explicit way and where they do they tend to use VASs (or non-preference-based valuation techniques). They are by implication invoking a set of assumptions more restrictive than the QALY model. The concerns raised in this section about preference-based measures apply *a fortiori* to health measures.

In conclusion, it is not possible to produce a definitive set of criteria for judging the appropriateness of the model of preferences. Instead, we suggest the researcher is explicit about the model of preferences and is asked to consider the likely validity of the underlying assumptions.

Valuation technique

Techniques used to value healthcare benefits include the VAS, ME, SG, TTO, PTO and WTP. Chapter 2 reported on the near-consensus in the economics literature that the VAS and ME do not have a basis in economic theory for estimating the strength of people's preferences and do not generate a measure suitable for use in CUA. Although there could be a case for the VAS, if it can be shown to be related to a choice-based method by a robust statistical model. This is reflected in the check-list. The issue of valuation techniques is fully explored in chapter 4.

Quality of data

All economic instruments for measuring healthcare benefits will use data elicited from valuation studies. These studies vary in terms of their respondents, the size of sample, and the method of administering the questionnaires (e.g. interview compared with self-complete administration, or with and without the aids of props). These have implications for the quality of the data in terms of the representativeness of the respondents, the reliability of the data, and the extent to which the respondent understood the task. The valuation of the larger generic health classifications also depends on there being a method of estimating values for all health states from the valuation of a sample of states (Dolan *et al.,* 1996). These aspects of valuation will now be examined on more detail:

- The background of the respondents should be examined and assessed to see if they are representative of the population whose values are being sought.
- Values should be reported with an indication of the reliability of the estimated values. Large variances should not be regarded as a fault of the measure since in part this may reflect genuine differences in preferences in the population. They might also be the result of a small sample size in the valuation survey. For the researcher, it is important to have access to the extent of the variation in order to conduct a sensitivity analysis. They may also wish to have access to a breakdown of the results by groups.
- Respondents' understanding of the task is partly reflected in the logical consistency of their answers. For some health classification systems and disease-specific descriptions it is possible to determine their ranking *a priori*. Where one 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 frequency with which this arises provides some indication of whether respondents understood the task. However, there are no accepted standards of consistency, and the analyst must judge whether respondents sufficiently understood the task to have confidence in the survey results. In some valuation surveys respondents displaying extreme cases of inconsistency are removed (Torrance *et al.,* 1982; MVH Group, 1994), but this may have implications for the representativeness of the sample.
- The response and completion rates of the valuation surveys should be reported since they have implications for all three aspects of quality. The rates may also affect the reliability of the data and the representativeness of the respondents, since there is a tendency for response rates to be lower/reduced in lower income groups. They may also indicate the respondents' difficulty with understanding the task and their acceptance of it.
- Some of the generic HSCs are too large for it to be possible to value all the health states directly. The 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. Only a sample of health states is valued for such instruments, and these are used to estimate values for all their states. There are two methods of undertaking this estimation (Froberg and Kane, 1989). One is by statistical inference, which involves the use

of multivariate techniques to estimate values for a functional form specifying the relationship between items of the HSC. The other is an algebraic approach, where individual utility functions are estimated for each dimension, and then aggregated using a function obtained by algebraic solution (Torrance, 1982). These raise substantial technical issues in their own right which are beyond the scope of this chapter.

Empirical validity – the acid test

The descriptive content of an instrument and the way it is valued provide a rationale for supposing whether or not a measure **could** generate values which reflect people's preferences. The ultimate test, of course, is whether the values do so in practice. There is very little evidence on the validity of instruments used in economic evaluation. However, given the importance of this issue we shall consider how this might be done. We propose a hierarchy of evidence, based around three types of evidence. The first is based on RP data, the second on stated preference data, and the third on hypothesised preferences.

Revealed preferences

The difficulties in obtaining RP data were raised earlier in this chapter. Experience from the literature on valuing life, such as by the use of labour market premiums for risky jobs, is plagued with problems of poor information and confounding (Mooney, 1977). However, it would seem to be worth developing this approach. There are some interesting developments testing stated WTP values (e.g. Chestnut *et al.,* 1996). An important point is that it does not require situations where the consumer has perfect information but it is necessary to know the consumers' **perception** of the benefits attributable to a product (which may not be health care) along with their purchasing decisions. There are many circumstances in health care, even in the UK, where people are buying health care out of pocket (e.g. complementary medicine and *in vitro* fertilisation. Even without payment people are making choices in health care. The problem is how to disentangle the patients' perceived benefits in order to assess preferences over the set of attributes. Such research is going to be difficult but should be a priority for the future.

Obtaining RP data from societal decision-making is equally fraught with problems, and these have been well reported in the literature, particularly in the context of valuing life There is again a problem of contaminants and confounding factors making it difficult to interpret the basis of the decisions made and hence preferences.

Stated preferences

Given the absence of RP data, an alternative to tests of the validity of a measure would be a comparison with stated preference.

A simple ordinal test would be to ask patients to rank health states they have experienced before and after surgery. This is a form of responsiveness that applies to those changes in health reported by patients rather than those deemed to be clinically significant. Testing the interval properties of a MAUS is more difficult. One approach would be to administer a direct method of preference elicitation on the same patient group and assess the convergent validity (e.g. directly elicited TTO values with those obtained from the Measurement and Valuation of Health (MVH) Group tariff for the EQ-5D). However, it is difficult to interpret differences since they may legitimately reflect the background characteristics of the respondents.

For testing the ability of a measure to predict social preferences, Nord (1991) has suggested a method based on the patient trade-off question. For example, if a value of 0.4 has been assigned to state A and 0.7 to state B, then this implies a subject is indifferent between making one patient in state A well for 2 years and making two patients in state B well for 2 years. This test also incorporates equity, and in this sense it also does not compare like with like. Furthermore, the validity of this technique as a means for obtaining social values has not been tested.

Hypothetical preferences

We also suggest a less direct method for testing validity based on hypothetical preferences, which examines whether the scores generated by an instrument reproduce the expected differences between groups of patients. This is a version of the psychometric test of construct validity, since the researcher must hypothesise or construct the expected differences (Streiner and Norman, 1989). It could be hypothesised, for example, that a patient would prefer a less severe condition, and hence this state should be associated with a higher score. This could be used to test the sensitivity of an instrument to expected differences between groups or its responsiveness to hypothesised changes (e.g. Hollingworth *et al.,* 1995; Katz *et al.,* 1994). The hypothesis must be chosen with some care, given the reservations already expressed about construct validity. We accept there are problems with this approach but believe that, used with care, it can provide useful insights into the validity of measures for use in economic evaluation.

A check-list for judging the merits of preference-based measures of health

The application of psychometric criteria of practicality, reliability and validity to measures of HRQoL has been reviewed and adapted for economic evaluation. The results are summarised in the form of a check-list for judging the merits of preference-based measures of health (*Box 3*). The criteria of practicality and reliability have been included, but not internal consistency. The tests of content, face and construct validity have been retained but limited in application to assessing the descriptive validity of the instrument. Up to this point in the check-list there is little disagreement with the criteria used to assess HRQoL measures. It is in the methods of valuation, and the empirical testing of validity of preferences, that there is a divergence, reflecting the different concepts being measured.

An important feature of the check-list is its comprehensiveness. It covers a larger range of characteristics than is usually discussed in the health economics literature, which has tended to focus on theoretical issues. This allows consideration to be given to the feasibility of using the instrument, its reliability (and hence the required sample size), and its validity in practice.

Where there is no apparent consensus between economists, the check-list requires the researcher to be explicit about the chosen method. The researcher is at least encouraged to ask questions about their chosen measure, such as the likelihood of the assumptions being violated and the existence of empirical evidence on validity.

The breadth of the check-list raises the inevitable question as to whether some parts of it are more important than others and hence should be given a larger weighting. It could be argued that it is important to consider whether an instrument is measuring the right concept first, that is validity, before the more practical concerns. To measure the right concept badly could be better than measuring the wrong one well. However, a theoretically superior measure which is not feasible to use has no practical value. Assessing the importance of different items in the check-list requires careful judgement on the part of the user.

We have chosen to call the result a check-list rather than a list of criteria, since the term criteria implies a degree of consensus that does not exist in the

BOX 3 Check-list for judging the merits of preference based measures of health

Practicality
- How long does the instrument take to complete?
- What is the response rate to the instrument?
- What is the rate of completion?

Reliability
- What is the test–retest reliability?
- What are the implications for sample size?
- What is the inter-rater reliability?
- What is the reliability between places of administration?

Validity

Description
- Content validity:
  - Does the instrument cover all dimensions of health of interest?
  - Do the items appear sensitive enough?
- Face validity:
  - Are the items relevant and appropriate for the population?
- Construct validity:
  - Can the unscored classification of the instrument detect known or expected differences or changes in health?

Valuation
- Whose values have been used?
- Assumptions about preferences:
  - What is the assumed model of preferences?
  - What are the main assumptions of this model?
  - How well are the preferences of the patients/general population/decision makers likely to conform to these assumptions (see examples in text)
- Technique of valuation:
  - Is it choice based?
  - Which choice-based method has been used?
- Quality of data:
  - Are the background characteristics of the respondents to the valuation survey representative of the population?
  - What was the degree of variation in the valuation survey?
  - Was there evidence of the respondents' understanding of the task?
  - What was the method of estimation? (where relevant)

Empirical
- Is there any evidence for the empirical validity of the instrument?
  - Revealed preferences?
  - Stated preferences?
  - Hypothesised preferences?

literature on what is a valid measure. At a recent presentation of an early draft of this chapter to the UK Health Economists Study Group meeting (Brazier and Deverill, 1997) there was a concern that it was too early to 'lay down the law'. The purpose of the check-list approach is to provide guidance rather than rules, and to acknowledge disagreement when it exists. The precedent in health economics is the widely used check-list for economic evaluation (Drummond *et al.*, 1987). The check-list presented here can be used to inform the design of an economic evaluation as well as in the review of instruments and published studies.

Conclusions

We have shown the inappropriateness of many psychometric criteria for judging the validity of measures for use in economic evaluation. We propose instead a check-list which incorporates the economists' perspective on the role of preferences. It will now be used to conduct a review of QALY instruments (chapter 5), the use of health status questionnaires in economic evaluations (chapter 6) and to review all economic evaluations published in 1995 using HSQs (chapter 7). Although the prime motivation for constructing the list was to help us conduct these reviews, we believe it could also be useful to researchers in the design and review of economic evaluations.

References

Arrow KJ. Uncertainty and the welfare economics of medical care. *Am Econ Rev* 1963;**53**:941–73.

Bowling A. Measuring health: a review of quality of life and measurement scales. Milton Keynes: Open University Press, 1991.

Brazier J, Deverill M. Criteria for judging preference based measures of health – learning from psychometrics. Paper presented to the Health Economists' Study Group Meeting, Liverpool, January 1997.

Buckingham K. A note on HYE (healthy years equivalent). *J Health Econ* 1993;**12**:301–9.

Carr-Hill RA, Morris J. Current practice in obtaining the "Q" in QALYs: a cautionary note. *BMJ* 1991;**303**:699–701.

Chestnut LG, Keller LR, Lambet WE, Rowe RD. Measuring heart patients willingness to pay for changes in angina symptoms. *Med Decis Making* 1996;**16**:65–77.

Cook J, Richardson J, Street A. A cost–utility analysis of treatment options for gallstone disease – methodological issues and results. *Health Econ* 1994;**3**:157–68.

Culyer AJ. The nature of the commodity health care and its efficient allocation. *Oxford Economic Papers* 1971;**24**:189–211.

Culyer AJ. Measuring health: lessons for Ontario. Toronto: University of Toronto Press, 1978.

Dolan P, Gudex, C, Kind P. Valuing health states: a comparison of methods. *J Health Econ* 1996;**2**:209–32.

Donaldson C, Gerard K. Economics of health care financing: the visible hand. London: Macmillan, 1993.

Donaldson C, Atkinson A, Bond J, Wright K. Should QALYs be programme-specific? *J Health Econ* 1988;**7**:239–57.

Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 1987.

Fitzpatrick R, Zeibland S, Jenkinson C, *et al.* A generic health status instrument in the assessment of rheumatoid arthritis. *Br J Rheum* 1992;**17**:439–47.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – I: measurement strategies. *J Clin Epidemiol* 1989a;**42**:345–54.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – II: scaling methods. *J Clin Epidemiol* 1989b;**42**:459–71.

Gafni A, Birch S. Preferences for outcomes in economic evaluation: an economic approach to addressing economic problems. *Soc Sci Med* 1995;**40**:767–76.

Gafni A, Birch S, Mehrez A. Economics, health and health economics – HYEs versus QALYs. *J Health Econ* 1993;**12**:325–39.

Guyatt GH, Sullivan MJ, Fallen EL, *et al.* How should we measure function in patients with chronic lung disease? *J Chronic Dis* 1985;**38**:517–24.

Hall J, Gerard K, Salkeld G, Richardson J. A cost utility analysis of mammography screening in Australia. *Soc Sci Med* 1992;**34**:993–1004.

Hollingworth W, Mackenzie R, Todd CJ, Dixon AK. Measuring changes in quality-of-life following magnetic-resonance-imaging of the knee – SF-36, EuroQol or Rosser index. *Q Life Res* 1995;**4**:325–34.

Hunt SM, McEwan J, McKenna SP. Measuring health status Beckenham: Croom Helm, 1986.

Hurst NP. A longitudinal study of patients with rheumatoid arthritis. Presentation to a Clinical Users Group of the EQ-5D, York, 1996.

Johannesson M, Jonsson B, Karlsson G. Outcome measurement in economic evaluation. *Health-Economics* 1996;**5**(4):279–96.

Katz JN, Phillips CB, Fossel AH, Liang MH. Stability and responsiveness of utility measures. *Medical Care* 1994;**32**(2):183–8.

Loomes G, Mckenzie L. The use of QALYs in health care decision making. *Soc Sci Med* 1989;**28**:299–308.

McDowell I, Newel C. Measuring health: a guide to rating scales and questionnaire. Oxford: Oxford University Press, 1987.

McHorney CA, Ware JR, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care* 1994;**32**:40–66.

Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and healthy-years equivalents. *Med Decis Making* 1989;**9**:142–9 (erratum: *Med Decis Making* 1990 Apr–Jun;**10**(2):148–9).

Mooney GH. The valuation of human life. London: Macmillan, 1977.

MVH Group. The measurement and valuation of health: first report on the main survey. University of York unpublished manuscript, 1994.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Plan Manag* 1991;**6**:234–42.

Nord E. Toward quality assurance in QALY calculations. *Int J Technol Assess Health Care* 1993;**9**:37–45.

Nunnally J. Psychometric theory, 2nd edn. New York: McGraw Hill, 1967.

Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994;**39**:7–21.

Richardson J, Hall J, Salkeld G. Cost–utility analysis: the compatibility of measurement techniques and the measurement of utility through time. In: C Selby Smith, editor. Economics and health: 1989. Proceedings of the 11th Australian Conference of Health Economists. Melbourne: Public Sector Management Institute, Monash University, 1989.

Sackett D, Torrance G. The utility of different health states as perceived by the general public. *J Chronic Dis* 1978;**31**:697–704.

Slevin ML, Stubbs L, Plant HJ, Wilsdon P, Gregory WM, *et al.* Attitudes to chemotherapy: comparing views of patients with cancer with those of doctors, nurses and the general public. *BMJ* 1990;**300**:1458–60.

Smith R, Dobson M. Measuring utility values for QALYs: two methodological issues. *Health Econ* 1993;**2**:349–55.

Stewart AL, Ware J, editors. Measuring functioning and well-being. Durham: Duke University Press, 1992.

Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press, 1989.

Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986;**5**:1–30.

Torrance GW, Boyle MH, Horwood SP. Applications of multi-attribute utility theory to measure social preferences for health states. *Operations Research* 1982;**30**:1043–69.

Ware JE, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care* 1992;**30**:473–83.

Wilkin D, Hallam L, Doggett MA. Measures of need and outcome for primary health care. Oxford: Oxford Medical Press, 1992.

Williams A. The role of the EuroQol instrument in QALY calculations. York: Centre for Health Economics, 1995.

Chapter 4

Review of the techniques of health state valuation

Introduction

As discussed in chapter 2, there are two components to estimating QALYs. The first involves describing the state or profile of a person's health; the second the valuation of these descriptions. This review focuses on the second of these components.

There are a number of techniques for the valuation of health states. The relative merits of these techniques has been a subject for debate for many years. There have been a number of informative reviews of the techniques (Torrance, 1986; Froberg and Kane, 1989; Richardson, 1994; Dolan *et al.,* 1996), and this review builds upon this earlier work, in particular using Froberg and Kane (1989) as a point of departure in the current literature.

We have focused on the specific techniques for the elicitation of preferences (or quality weights) for use in the valuation of health states. We have not concerned ourselves with the issues surrounding whose values to elicit and in a wider sense the issues relating to framing and context effects, other than those which appear to be specific to particular valuation techniques. We regard these topics as general QALY issues, rather than method-specific concerns. We review those techniques which have been used to elicit health state utility values on a scale of 1 or less (as discussed in chapter 2), and consider these to be measures of strength of preference (i.e. to have cardinal scale properties). Therefore, this review does not include the literature covering contingent valuation or WTP techniques (which deserve attention in their own right). The health state valuation techniques covered in this review are the SG, TTO, the VAS (RS), ME and PTO.

We review the literature relating to these techniques and report on both the methodological and empirical findings to provide an assessment of the relative merits of the techniques.

We begin by providing a description of the health state valuation techniques reviewed, and details of the search methodology and search results. We then present the criteria used to review the performance of the techniques and a separate review of each of the techniques, followed by a comparison of the techniques. A discussion of the literature covering the relationships between the techniques follows, and conclusions are presented. An empirical listing of reported health state valuation studies and a listing of the literature reviewed are contained in the appendices to this report (see appendix 2).

Description of health state valuation techniques

Visual analogue scale

> A typical rating scale consists of a line on a page with clearly defined end-points. 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)

The VAS, sometimes referred to in the literature as the category rating (CR) scale or just the rating scale (RS) is simply a line, usually with well-defined end-points, on which respondents are able to indicate their preferences. The example shown in *Figure 1* is the 'thermometer rating scale', used by the EuroQol Group, which has the best imaginable health state at the top of the line and worst imaginable health state at the bottom. Proponents claim the VAS technique generates an interval scale measure of preferences, so that the differences in a person's strength of preference for a move between 90 and 95 on the scale should be the same as between 20 and 25 (Kaplan *et al.*, 1979).

This form of scaling was originally developed in psychophysics to measure people's response to sensory stimulation, such as light, sound and heat. It was developed in the field of psychometrics to assess feelings and attitudes across a range of fields of enquiry. It has been widely used in health research to measure health status, including people's perception of their symptoms, such as

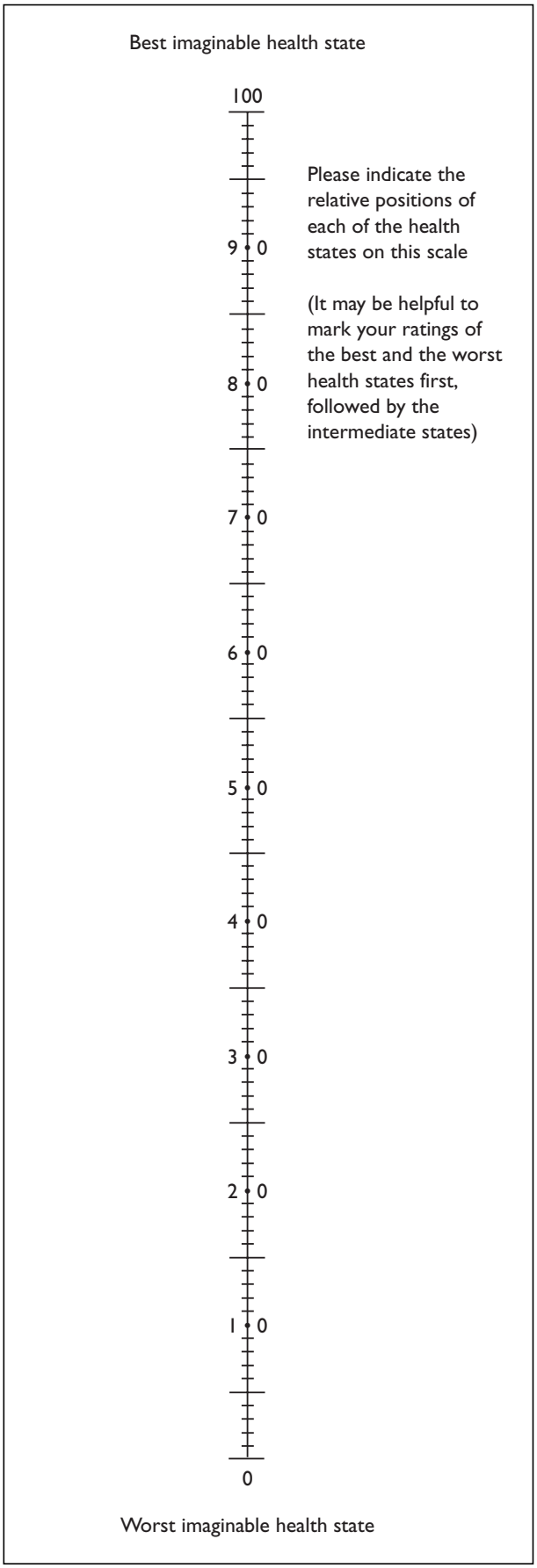


FIGURE 1 *VAS used by the EuroQol Group*

pain, functioning, and mental well-being (e.g. Nicholl *et al.*, 1992), where the categories are sometimes described verbally to assist the respondent. It has been extensively used to assess peoples valuation of different states of health.

There are many variants of the technique. The lines can vary in length, be vertical or horizontal and may or may not have intervals marked out with different numbers. The QWB MAUS, for example, was valued by asking respondents to place health states into one of 15 numbered slots, where 0 was death and 1 was optimum health. The EQ-5D has been valued using the finer interval markings of 0 to 100, reproduced in *Figure 1*. For some applications, respondents are asked to value a set of hypothetical health states on the same scale. They may be asked to place the best and worst at the endpoints as described by Torrance above, but this is not the case for the EuroQol version where the respondent is free to place the states on the scale in any order. Torrance and colleagues have also developed a felt board on which the respondent places cards describing different health states, whereas the EuroQol version asks respondents to indicate the position of a health state by drawing a line on a piece of paper on to the scale.

The VAS has been widely used to value health states, including all the MAUSs. The QWB has only been valued by the VAS, and the HUI-II and HUI-III have been valued by transforming VAS values into SG. It is therefore an important technique to review.

Where we refer to the VAS we use the term to capture all the general techniques which are characterised by the methodology i.e. RSs, CR and VAS methods. Where particular authors have referred to descriptions such as RS or CR we have used the general description of the VAS to describe them.

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 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)

ME was developed in psychometrics to measure sensory and non-sensory perception as an alternative to the VAS (Stevens, 1966). The phrasing of the question was intended to generate data with

ratio properties (i.e. one state is so much better or worse than another) and therefore it is often referred to in the psychometric literature as ratio scaling.

In the original valuation of the Rosser disability/distress classification, respondents were asked to rank and value six marker states, and then to value five of these against the least ill state (Kind *et al.*, 1982). All the remaining 23 states were valued against the marker states, including death. It was then possible to transform the value of all states on to a full health and death scale. There have been important variations in the versions used to value health states. Rosser and Kind (1978) report asking respondents to indicate how undesirable one state was compared to another, whereas Kaplan *et al.* (1979) asked how many times more desirable one state was compared to another. The version used by Sintonen (1981) provides the respondent with a scale from 0 to 100 for answering the question.

The main applications of ME to the valuation of health states includes the Rosser classification and the 15D. It has also been examined by Patrick *et al.* (1973) and Kaplan *et al.* (1979) as an alternative to the VAS for the QWB, but they claimed it was inferior to the VAS (see arguments below). It has not been widely used to value condition-specific health states.

Standard gamble

Referring to *Figure 2*:

> 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)

The respondent is asked to make a choice between alternative outcomes, where one of them involves uncertainty. They are asked how much in terms of risk of death, or some other outcome worse than the one being valued, they are prepared to accept in order to avoid the certainty of the health state being valued. This technique is based on the EUT of decision-making under uncertainty developed by Von Neuman and Morgenstern (1944). This theory rests on a set of axioms about the nature of individual preferences over uncertain prospects (see chapter 2). According to this theory the probability of success at which they are indifferent between the alternatives, which means they find them equally desirable, provides a unit of measurement for the value of the certain health state. The validity of the axioms and hence the SG technique is examined below.

There are many versions of the SG technique. The problem of explaining probabilities to respondents has been addressed through the use of various visual aids, such as the probability wheel developed by Torrance *et al.* (1976, 1986). Rather than asking the respondent an open question on their point of indifference, they are helped in arriving at some point of indifference by iterating between values for the probability of success p towards a point of indifference (i.e. the 'ping-pong' method). An alternative variant has been developed by Jones-Lee and colleagues (1993) without the use of a visual aid. Instead they have developed 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 are confident they would reject treatment. Finally, they are asked to indicate the value where they find it most difficult to choose.

SG has been modified to value states worse than death (Torrance, 1986). Another variation of SG is to use different reference or anchor states in

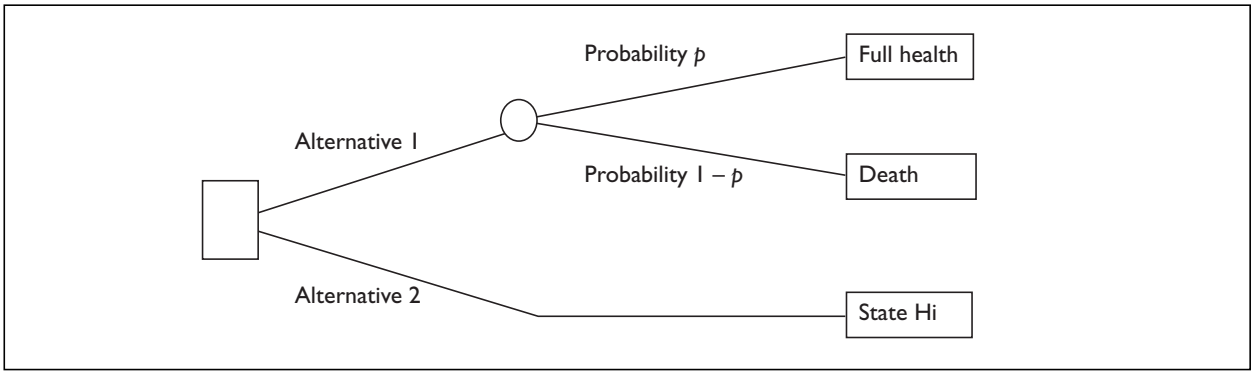


FIGURE 2 *Standard gamble*

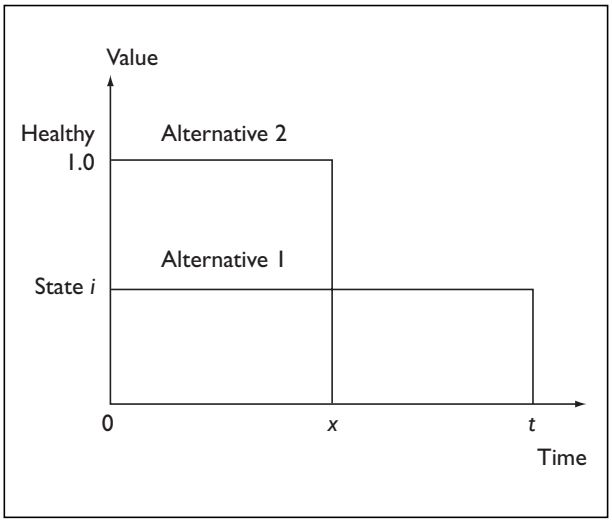


FIGURE 3 *Time trade-off*

alternative A. This can be useful where the state being valued is comparatively mild or temporary, and most respondents would be unwilling to contemplate a risk of death in the range being considered in the question. The values derived for health states from such gambles can be 'chained' back to the full health–death scale, provided the reference state is valued against full health and death in another gamble.

The technique has been widely used in the decision-making literature (Keeney and Raiffa, 1976). It has been extensively applied to medical decision-making, including the valuation of health states, where it has been used (indirectly via a transformation of the VAS) to value the HUI-II and HUI-III, and also to condition-specific health state vignettes.

Time trade-off

Referring to *Figure 3*:

> 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)*

The TTO technique was developed by Torrance *et al.* (1972) as an alternative to SG, designed to overcome the problems of explaining probabilities to respondents. The respondent is asked to choose between two alternatives, both with certain prospects, that is, years in full health (x) and years in the health state being valued state (t). The respondent is directly asked to consider trading a health improvement for a reduction in their length of life. The health state valuation is the fraction of healthy years equivalent to a year in a given health state, that is, x/t.

Visual aids have been developed to assist the respondent, and again Torrance utilises a 'ping-pong' style for eliciting preferences. He has also developed a version for valuing states worse than death. For very mild or temporary states, where the respondent may be unwilling to consider trading survival on the scale being offered them, Torrance has developed methods of chaining from questions by replacing death with a poor state of health and indirectly deriving a value on a full health death scale by a process similar to the one described for SG.

TTO has been tailor made to value health states, and it has been used extensively for this purpose. It was administered to value the first version of the HUI-I and in a large UK survey to value the EQ-5D, as well as numerous condition-specific health states.

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)

The PTO technique is a way of estimating the social value of different health states. As with TTO and SG, the PTO technique asks the respondent to make a choice between alternatives. The crucial difference is that the respondent is asked to make a choice in the context of a decision involving other people rather than themselves. PTO

* 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 $X/(t - x)$.

basically consists of asking people how many outcomes of one kind (e.g. outcome A) they consider to be equivalent in social value to *x* outcomes of another kind (e.g. outcome B) (Nord, 1995). The trade-off is between one group of people experiencing one gain against people experiencing another gain. The context of this social choice has attracted the interest of those economists, most notably Nord (1992), who regards this as more relevant for social choice contexts than the conventional individual perspective of the other valuation techniques.

This technique was originally known as the equivalence technique (Patrick *et al.*, 1973; Torrance, 1986), but has been renamed by Nord as the PTO method to better reflect the choice being presented to the respondent. The variants of the technique involve changing the reference group and changing the presentation of the question. To generate the social value of any medical value in numerical terms, Nord has proposed using a standard reference gain against which all other gains can be measured. He has proposed using the saved young life equivalent (SAVE), which reflects saving a young life and returning the patient to full health (Nord, 1992).

Although PTO values may be viewed as social values, as respondents are making choices between treating different groups of patients, the focus of such choices, on an individual basis, can be influenced by many characteristics. For example, PTO responses can be a function of health status before intervention (i.e. initial severity), health status after intervention, the size of the health gain offered by the intervention or whether patients are receiving a life saving or life improving treatment or a combination of many such issues.

This technique has not been used widely to value health states. We have used the PTO notation to describe the technique, including applications referring to the equivalence technique.

Details of the literature search methodology

Structure of the literature search process

The search strategy was developed via an iterative method, with an information specialist and a health economist working in tandem, whereby preliminary search results were sought and reviewed as a means of 'brainstorming' the numerous different variants available.

Preliminary work, as described elsewhere (see chapter 3), had highlighted the deficiencies of using indexer-assigned subject terms for the search strategy. In fact, using MEDLINE as an example, only the terms 'HEALTH-STATUS-MEASUREMENT' and 'QUALITY-OF-LIFE' reflect any of the facets of the review with no corresponding terms to convey the idea of preference measures. It was clear from this that what was required was free-text searching (i.e. of titles, abstracts, etc.) combining as many variants and permutations as deemed possible.

The following search terms were identified to capture literature relating to valuation techniques (using a MEDLINE exemplar):

(1) (RATING SCALE*) or (CATEGOR* near2 SCAL*) or (LINEAR SCAL*) or (LINEAR ANALOG*) or (VISUAL ANALOG*)
(2) (MAGNITUDE ESTIMATION) or (RATIO SCAL*)
(3) STANDARD GAMBLE*
(4) (TIMETRADEOFF) or (TIME TRADEOFF) or (TIME TRADE OFF) or (TIME TRADE*)
(5) (PERSONTRADEOFF) or (PERSON TRADEOFF) or (PERSON TRADE OFF) or (PERSON TRADE*) or (EQUIVALEN* near2 NUMBER*)

Due to the limited literature and the specific nature of the search terms, searches 3 and 4 above (SG and TTO) were applied directly to data sources. Given the size of the literature relating to searches 1, 2 and 5, it was necessary to improve the specificity of retrieval for these terms to identify only those articles that referred to health state valuation and/or preference measurement. The preferred way of operationalising the search was to construct a results set related to these terms. The following subset was developed:

(HEALTH near2 STATE*) or (HEALTH near2 STATUS) or HEALTH-STATUS* or (HEALTH near2 UTILIT*) or (QUALITY near2 LIFE) or QUALITY-OF-LIFE*

With reference to this subset, and the above search terms 1 to 5, it will be noted that the non-standardised use of terminology necessitated the frequent use of the proximity operator 'near2', that is, the two target words occurring within two words of each other. As can be seen from the search results (below), the emphasis of the search strategy was generally on sensitivity at the expense of specificity.

All searches covered the period from the date of commencement of each database service (e.g. MEDLINE from 1966) to November 1997 but were restricted to English language materials.

Databases used

The nature of the topic necessitated the searching of health-related databases such as MEDLINE, EMBASE (*Excerpta Medica*), HealthSTAR and the Citation Index for Nursing and Allied Health and Sociofile (CINAHL), general science and social sciences databases such as the Science Citation Index, the Social Science Citation Index and the International Bibliography of the Social Sciences (via the BIDS service), and an economic-specific database, EconLit. The NEED from the NHS Centre for Reviews and Dissemination was also searched but, with its emphasis on applications rather than methodology, this proved of little value. All citations resulting from the search strategy were retrieved and articles were reviewed for relevance on the basis of their abstracts. Material from these databases were supplemented with relevant articles from the authors personal knowledge and experience.

Search results

From these searches a total of 1581 articles were initially identified, see *Figure 4* for further detail. The identified abstracts and bibliographic details were reviewed for relevance by a health economist and photocopies of relevant items were requested. In cases of doubt the article in question was obtained and a subsequent judgement on relevance made based on the full article.

The identified abstracts and bibliographic details were reviewed according to the following criteria:

(1) References had to present discussion/results relating to at least one of the designated health state valuation techniques.
(2) References had to discuss/present health state valuation techniques in the context of health and health care evaluation, that is, literature relating to environmental or transport applications, for example, were not selected for review.
(3) The valuation technique(s) had to be discussed/applied in the context of eliciting values for general multi-dimensional health

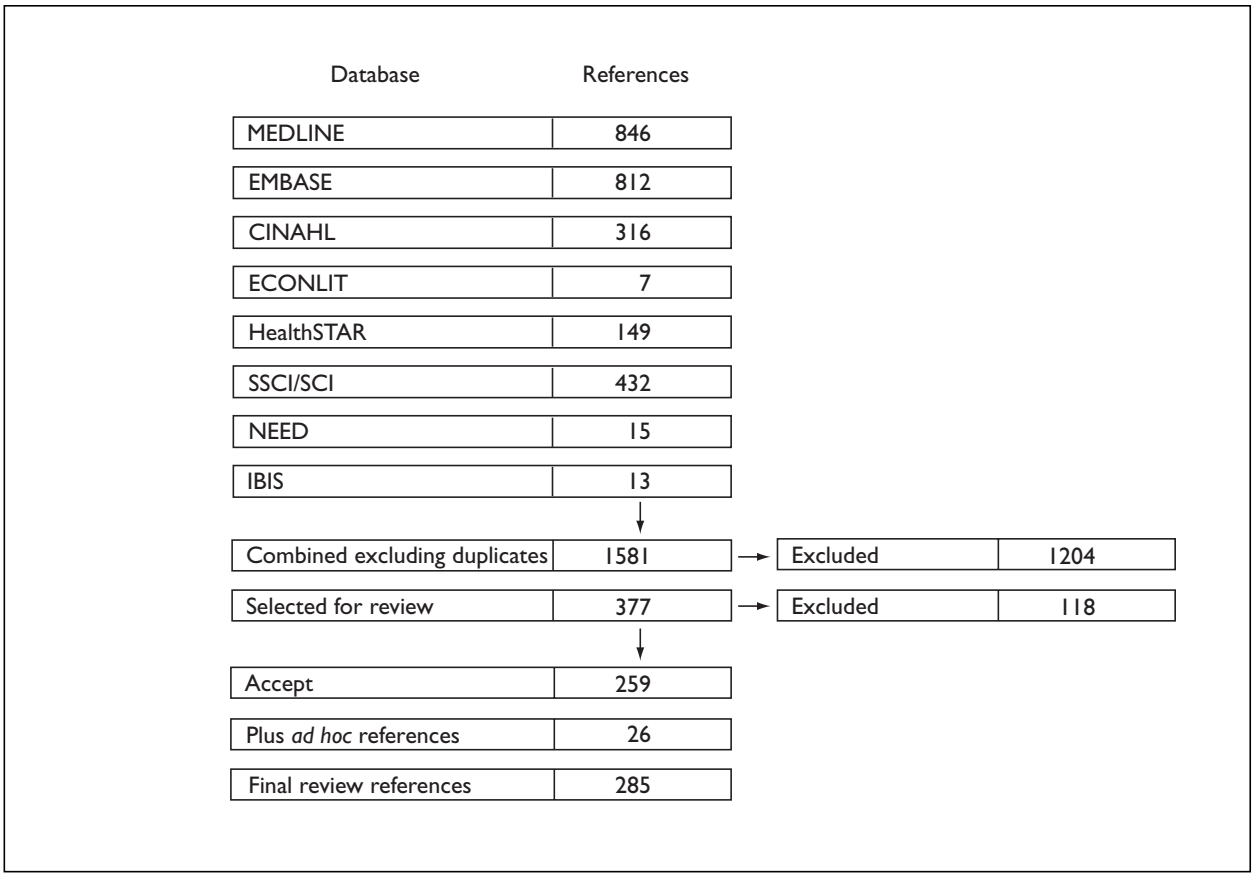


FIGURE 4 *Overview of the literature identification process*

states, that is, where techniques had been discussed/applied in the context of valuing unidimensional health state descriptors (e.g. pain **or** disability only) they were not selected for review.

Applying these criteria at the initial sifting stage a significant number of references were excluded (1204 at this stage). A large number of the references that were excluded had been identified using search terms related to VAS techniques, and on inspection they were studies using the technique to measure individual dimensions of health status e.g. pain.

The resultant set of 377 articles were obtained and once again the above criteria were applied. As a result of this a further 118 articles were rejected; again these articles were mainly VAS applications other than for the purposes of general health state valuation. A further 26 references were identified from bibliographic information contained in the identified literature, and these were classed as *ad hoc* references.

Including *ad hoc* references we found a literature of 285 articles for review.

Criteria for reviewing performance

The basic concepts of practicality, reliability and validity form the criteria used for reviewing the performance of the valuation techniques. The check-list developed in chapter 3 for judging preference-based health questionnaires against these criteria required some adaptation for this specific review. The methods for testing practicality and reliability are the same as those presented. However, the concern with descriptive validity, in terms of how well the HSC system describes health is plainly not relevant, while the notions of theoretical and empirical validity must be extended to deal with some of the nuances of the differences which arise between the valuation techniques.

The criteria developed to undertake the review are described below.

Practicality

The practicality of an instrument depends on its acceptability to respondents. 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).

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 retest reliability, or between raters. All measures have some degree of random variation, and the consequences of more random variation is the need for a larger sample size. A more serious problem arises if there is evidence of a systematic difference in health state values, such as an increase in health state values with repeated administrations. Correlation coefficients are the commonly presented measures of reliability. Whilst we recognise the arguments against the use of correlation as a measure of agreement between measures (Bland and Altman, 1986), most studies have only reported on reliability using this type of summary statistic.

Theoretical validity

The basis of the techniques in economic theory has been argued by some economists to be paramount in the assessment of the validity of a measure to be used in economic evaluation (Gafni, 1996). The debate about the importance of theory *per se* has already been examined in chapter 3. This review examines the validity of the theoretical basis used to support the valuation techniques (e.g. SG and EUT).

The assessment of the theoretical basis of the techniques also helps identify the testable assumptions which underly the techniques. These provide a means of testing the empirical validity of the techniques.

Empirical validity

As discussed in chapter 3, the 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. In the absence of RP data, some economists have been dismissive of even trying. However, our review of the literature has uncovered numerous attempts to examine empirical validity and these are reviewed here. These include evidence on the assumptions underlying the different techniques, and tests of the values generated by the techniques against stated and hypothetical preferences. These issues are considered in further detail below.

Testing the theoretical basis of the techniques

The review of theoretical validity identifies the assumptions and testable predictions of the theory underlying the different techniques of valuation. These include, for example, whether or not peoples' attitude to risk is constant (SG), or whether they have a zero time preference (TTO). Evidence on the extent to which theory correctly describes the individual preferences provides a means of empirically testing the techniques.

However, these tests assume that the 'acid test' is the descriptive accuracy of the underlying theory. Others have suggested that the basis of the techniques is that they offer a more rational basis for decision-making. This is a form of 'normative' validity, and requires that the assumptions have some kind of normative appeal to individuals or the decision-makers concerned (Gafni, 1996). Consideration of this is more difficult since it has not been tested.

Testing the techniques against stated preferences

The empirical validity of the values generated by the techniques can be tested against other measures of stated preferences. These could be stated ordinal preferences. For example, when values are directly elicited from patients before and after a trial, do the results confirm the directly elicited views of patients concerning whether or not they prefer the after treatment state to the one pertaining before? Such an approach has not been used in any study found in this review. There are studies where respondents are asked to rank hypothetical health states and to value them using different techniques. The ability of each technique to correctly predict the ranking of the states provides some evidence of their ordinal properties.

The valuations elicited by the techniques should also possess interval properties for use in economic evaluation. The only method for examining this property has been to assess the degree of convergence between the different valuation techniques, but this cannot provide conclusive proof (see the argument in chapter 3).

Testing the techniques against hypothetical preferences

Researchers may hypothesise that one health state should be preferred to another. For example it could be hypothesised on the basis of past experience that patients with renal failure would prefer to be in a health state following a successful transplant than to depend on dialysis. Another test has been to examine the extent to which health state valuations for MAUSs are consistent with the scale. For many pairs of health states defined by a MAUS classification such as the EQ-5D, one state can be regarded as dominant over the other if it is less severe on one health dimension or more and no worse on the remaining dimensions. It is hypothesised that the dominant health state should be logically preferred or regarded as equal to the other state. The degree of logical consistency can be examined in terms of the classification of strict consistency with the predetermined rank (i.e. >), strict inconsistency in the case of reversals (i.e. <), and equality (i.e. =). These tests can be undertaken at the aggregate level, that is, on the summed responses of respondents, but a more rigorous check on respondent understanding is performed on each individual's answer (e.g. see MVH data: Gudex *et al.*, 1996 and Dolan *et al.*, 1996a). High levels of inconsistency could reflect confusion on the part of the respondent with the valuation task. Comparisons between techniques can only be undertaken where the valuation techniques have been used on the same MAUS classification and the same set of respondents.

Review of health state valuation techniques

SG review

Practicality

Empirical studies using SG have reported response and completion rates to demonstrate an acceptable performance in terms of practicality. Many studies, across different respondent groups, have reported completion rates between 95 and 100% (Rabin *et al.*, 1993; Patrick *et al.*, 1994; Morss *et al.*, 1994; Ramsey *et al.*, 1995; Dolan *et al.*, 1996b; Lenert *et al.*, 1997). Studies by Revicki (1992) and Gage *et al.* (1996) report completion rates over 80%. Whilst many studies report that SG has proved to be feasible they do not report quantitative details (e.g. Llewellyn-Thomas *et al.*, 1982). Clinically based studies have found SG to be feasible and acceptable amongst patient groups, for example, cancer patients (Llewellyn-Thomas *et al.*, 1982) and lung transplant patients (Ramsey *et al.*, 1995).

Although SG has shown some completion problems within particular studies, these have been no worse than similar difficulties associated with other instruments used at the same time. For example, where completion problems

occurred in studies by Patrick *et al.* (1994) and van der Donk *et al.* (1995), which used SG, TTO and VAS methods, SG was not seen to be more burdensome than other methods employed. Some studies report completion problems with SG (Revicki, 1992; Stiggelbout *et al.*, 1994), and Stiggelbout *et al.* comment that questions were too hypothetical, but no general pattern emerges from the literature.

Empirical studies are predominantly interview based, with some self-completed tasks. We have not identified evidence to demonstrate the practicality of SG in postal questionnaire format. Bosch and Hunink (1996) report the use of SG via a postal formats combined with a telephone interview, finding an acceptable completion rate (11% refused to answer the SG and TTO questions). Although some commentators report SG as a complex method of valuation (e.g. Froberg and Kane, 1989), empirical findings suggest that the SG can be an acceptable method of health state valuation provided due care is taken with its administration.

Reliability

Froberg and Kane (1989) present evidence of good intrarater reliability ($r = 0.77$; from Torrance, 1976) and test–retest reliability ($r = 0.80$). *Table 1* details more recent studies, adding to the review undertaken by Froberg and Kane, which provide further support for the reliability of SG (Boyd *et al.*, 1990; Reed *et al.*, 1993; Bakker *et al.*, 1994; O'Brien and Viramontes, 1994; Gage *et al.*, 1996; Dolan *et al.*, 1996b).

TABLE 1 *Test–retest reliability of the SG, TTO and VAS techniques*

Test–retest reliability	SG	TTO	VAS
1 week or less	0.80[a] 0.77–0.79[j]	0.87[a]	0.77[a] 0.70–0.95[j]
4 weeks	0.82[b]	0.81[c] 0.63[d]	0.62[b] 0.89[d]
3–6 weeks		0.50–0.75[l]	
6 weeks		0.63–0.80[c] 0.85[e]	
10 weeks		0.73[f]	0.78[g]
6–16 weeks	0.63 – props[m] 0.74 – no props[m]	0.83 – props[m] 0.55 – no props[m]	
1 year	0.53[h]	0.62[h]	0.49[h]
Other (time unspecified)	0.82[i] 0.80[k]	0.74[i] 0.67–0.92[k]	

Correlation's undertaken where specified: intraclass correlation coefficient – b, f, g, j, k; Pearson correlation coefficient – d, m; others unspecified

Note: see appendix 2 for qualitative comments concerning the reliability of ME and PTO

[a] *O'Connor* et al. *(1985),* [b] *O'Brien and Viramontes (1994),* [c] *Churchill* et al. *(1987),* [d] *Gabriel* et al. *(1994),* [e] *Molzahn (1996),* [f] *Dolan* et al. *(1996a),* [g] *Gudex* et al. *(1996),* [h] *Torrance* et al. *(1976),* [i] *Reed* et al. *(1993),* [j] *Bakker* et al. *(1994),* [k] *Gage* et al. *(1996) (range 1.3 to 28 weeks),* [l] *Ashby* et al. *(1996),* [m] *Dolan* et al. *(1996b)*

Theoretical validity

As discussed earlier in this chapter and in chapter 2, SG is based on EUT. It is the most widely used model of behaviour under uncertainty, and for many years it has held a dominant position, as an explanation of choice under uncertainty, both in the teaching and the application of economics. Due to its theoretical underpinnings, SG is viewed as the classic method of decision-making under uncertainty (Gafni, 1994) and due to the uncertain nature of medical decision-making SG is frequently referred to as the criterion or reference method of health state valuation (Boyd *et al.*, 1990; Llewellyn-Thomas *et al.*, 1996; Kavanagh *et al.*, 1996), and often classed as the 'gold standard' (Torrance, 1996; Gafni, 1994). Such theoretical support has led to the automatic use of SG methods on these grounds, possibly without due consideration given to other methods or to study specific characteristics. For example, in a study by Nichol *et al.* (1996) the authors clearly state that 'because it [SG] is considered to be the criterion method for eliciting these evaluations [patient utilities] we elected to use the SG to assess patients' utilities'. If an individual behaves in agreement with the axioms of EUT, then the SG method is said to yield utility values. Yet, if the axioms are violated, as is increasingly suspected, the justification for preferring the SG method over alternative valuation techniques, on the basis of theoretical strength, must come into question.

Further to the arguments surrounding EUT is the debate concerning the use of risk and probabilities (as the unit of measure by which utilities are elicited), and the related issue of attitudes to risk, especially risk aversion. As medical decisions usually involve uncertainty the use of the SG method would seem to have great appeal. However, the version of SG used to value

health states does not value uncertainty but uses risk to value a certain state (Richardson, 1994; Broome, 1993; Buckingham, 1993). There is concern surrounding the influence of factors in the decision other than the preference for a health state, such as a gambling effect or risk aversion (Richardson, 1994; Broome, 1993). Richardson (1994) highlights that in using risk as a unit of measure, in the elicitation of health state preferences, such preferences are exposed to the effects of a 'specific utility of gambling' or a 'specific utility of risk' arising from risk *per se,* and points out that 'Von Neuman and Morgenstern did not believe that their axioms accounted for the specific utility of risk'. Such an effect gives rise to concern over the use of the SG method as it could introduce an 'additional, random element whose relationship to the specific utility of the risk associated with a medical procedure is unknown' (Richardson, 1994).

The possible influence of attitudes to risk within the SG has raised some theoretical concerns (Kahneman and Tversky, 1979). A respondent's attitude to risk may be risk averse, risk neutral or risk seeking, and at times may be a mixture of all three (Loomes and McKenzie, 1989). Individuals attitudes to risk are liable to affect the choices they make between different alternatives, therefore, SG valuations may be affected by differing attitudes to risk.

The theory underpinning the SG is undoubtedly theoretically appealing and it has its prominent supporters, nevertheless, there are strong arguments against the theory.

Buckingham *et al.* (1996) reject the idea, implied by authors such as Torrance and Mehrez and Gafni, that SG provides a 'gold standard', stating that too much doubt has been cast on the validity of its underlying axioms (by authors such as Loomes and Sugden (1982) and Kahneman and Tversky (1979)) to support such a supposition. Richardson (1994) presents strong theoretical arguments against the interpretation of the Von Neuman and Morgenstern theory and has dismissed SG as the gold standard, stating that the foundations of SG in economic theory cannot be accepted as the basis of measurement in CUA. Johannesson (1994) presents descriptive evidence to show that individual decisions often violate EUT and Wakker and Stiggelbout (1995) state that 'critical tests of EUT in decision-theory literature have shown that EUT is not empirically valid'.

Empirical validity

Evidence: theory. The theoretical underpinnings of the SG (i.e. EUT) have been subject to growing scrutiny. For example, according to EUT, the utility assigned to a particular health state should not be influenced by the alternative outcomes offered in the gamble (i.e. the independence axiom), rather respondents are supposed to adjust their indifference probability to allow for alterations in the gamble outcomes. Llewellyn-Thomas (1982) report an empirical study to examine this aspect of EUT.* The authors found that SG utilities were strongly influenced by the characteristics of the 'failure' outcome of the gamble, a violation of EUT. Gage *et al.* (1996) also found that SG utilities are influenced by the outcome of the gambles. Further evidence that respondents systematically violate the axioms of EUT in the health context has been further reported by Dolan *et al.* (1996) and Read *et al.* (1984) and more generally by Hershey (1981) and Schoemaker (1982).

Evidence highlighting differing attitudes to risk has raised further concern over the validity of EUT. Individuals have been shown to exhibit differing attitudes to risk (Bakker *et al.*, 1994; Wakker and Stiggelbout, 1995; Rutten van Molken *et al.*, 1995b; Clarke *et al.*, 1997). Kahneman and Tversky (1979) have argued that respondents generally act as if they are risk averse when choices are framed in terms of potential gains and as risk seeking when choices are framed in terms of potential losses. Loomes and McKenzie (1989) reviewed evidence that individuals exhibit a mixture of risk aversion and risk seeking and that an individuals attitude to risk cannot be represented by a constant value. Kahneman and Tversky (1982) also find that individuals tend to overestimate small probabilities and underestimate large probabilities and suggest that probabilities of less than 0.1 and greater than 0.9 present individuals with difficulties, which subsequently raise concerns in health state valuation tasks. Wakker and Stiggelbout (1995) discuss the possibility that people do not treat probabilities in a linear manner, as EUT supposes, but that people transform probabilities into decision weights, with probability transformation usually in an 'S' shape, where small probabilities are overestimated and large probabilities are underestimated. Some of the alternatives to EUT,

* Llewellyn-Thomas *et al.* varied reference states in the SG task and then 'chained' back to the death and full health scale; they then compared SG values to those directly obtained in a SG using death and full health as the outcomes.

such as Kahneman and Tversky's prospect theory* may help to explain some of the inconsistencies in responses to SG questions, but also serve to highlight the unstable nature of attitudes to risk. The evidence relating to the variability of individuals' attitudes to risk are not compatible with the axioms of the EUT.

Llewellyn-Thomas *et al.* (1996) discuss SG in the context of the criterion method to obtain utilities and state that 'because people's decision behaviours often are not congruent with the axioms of rational choice, the validity of using this prescriptive method [EUT] to describe an individual's actual decision-making, or to select the 'best' treatment strategy for that individual, has to be challenged' (Llewellyn-Thomas *et al.* cite Schoemaker (1982) and Fischoff *et al.* (1988)).

Evidence: preferences. Assessing the performance of SG in the context of stated preferences we have considered the convergent validity of SG and other valuation techniques, finding evidence to support a reasonable relationship between SG and TTO values (Torrance, 1986; Dolan *et al.*, 1996b; Bosch *et al.*, 1996; Reed *et al.*, 1993; Zug *et al.*, 1995), discussed further under TTO. Further tentative evidence of validity is offered by Dolan *et al.* (1996b) who report that all methods (SG and TTO both with props and without) produced a similar ordinal ranking of health states. Evidence was also found which indicates that SG does not correlate well with either health status (see review in chapter 6) or VAS methods (Rutten van Molken *et al.*, 1995b; Bakker *et al.*, 1995). SG values have generally been found to exceed those of VAS methods (Zug *et al.*, 1995; Gage *et al.*, 1996; Morss *et al.*, 1994; Revicki *et al.*, 1996; O'Brien *et al.*, 1994; Bakker *et al.*, 1995; Van der Donk *et al.*, 1995). Such empirical evidence would suggest that SG utilities are something other than a measure of health status.

With respect to hypothesised preferences we have found empirical evidence relating to the consistency of SG responses with expected rankings. Dolan *et al.* (1996b) examine the performance of SG (and TTO) against 12 logically consistent comparisons (EQ-5D health states) and report that SG produced high levels of consistency (population sample of 335 respondents). They report consistency rates of 83.8% for SG with props and 87.5% for SG with no props. Rutten van Molken *et al.* (1995b) consider SG responses against hypothesised preferences based on the natural underlying order of health state descriptions used (fibromyalgia patients, n = 85; anklyosing spondylitis patients, n = 144), and report a good level of consistency in SG responses. Although they report that 21% of patients completing SG provided at least one inconsistent response, about 70% of the inconsistent responses had differences which fell within the bounds of the standard error of measurement. Llewellyn-Thomas *et al.* (1982) report that SG provided a high level of consistency against an expected rank ordering of health states (cancer patients, n = 64). They report 54 (84%) of 64 respondents ranked five health states via SG in accordance with *a priori* expectations; with nine of the other respondents providing a rank order with only one inconsistent pairing (at interview, one).

Dolan *et al.* (1996b) assessed the empirical validity of SG responses (and TTO) on the basis of *a priori* expectations, developed from the results of previous studies. The authors hypothesised that valuations would not differ according to the age, gender, and employment status of respondent and that higher valuations would result for respondents with experience of illness. Results supported the hypothesis developed; however, the insight into the empirical validity of SG (and TTO) is tempered by the contentious nature of some of the hypothesised preferences applied.

One concern with SG is the potential 'unwillingness' of individuals to accept any level of risk (not even a very small risk) to obtain an improvement in health. Choice-based valuation techniques such as SG and TTO rely on the willingness of respondents to trade risk of death or life-years in order to improve their state of health. Consumer theory assumes individuals will trade to maximise utility. A reluctance to trade-off may be related to a number of issues, such as the context of the choice (e.g. Kahneman and Tversky, 1979), or to a misunderstanding of the task. With SG it may be that extreme risk aversion creates an unwillingness to take on risk. Such explanations have not been investigated in the current literature and need further attention (although Scott (1998), in

* Another is regret theory (Loomes and Sugden, 1982), where the value a person assigns to a health state depends not only on that health state but also on how that health state compares with the health state the person might have had if he or she had made a different choice. Subjects may shy away from the gamble choice in SG due to regret aversion. Regret may occur if they 'lose' the gamble and end up with the worst outcome.

a conference paper, begins to address the issue). A number of authors have reported results indicating an unwillingness to take on risk amongst some respondents (Stiggelbout *et al.*, 1994; Reed *et al.*, 1993). Reed *et al.* (1993), for example, report that of 35 patients completing SG and TTO valuation tasks, 16 were unwilling to accept any significant risk of death, whilst only nine of these 16 gave a similar response to the TTO task (i.e. zero trade-off). Although, findings of this nature are not widely reported in SG results, a number of studies report aggregate values very close to one for particular health states, without reporting frequencies (e.g. Bass *et al.*, 1996; Gage *et al.*, 1996; Dolan *et al.*, 1996b).

VAS review

Practicality

Although there are studies which report problems with response and completion (Brooks *et al.*, 1991), VAS methods have been widely accepted as the most feasible and acceptable of the health state valuation techniques, demonstrating high response rates and high levels of completion (Torrance, 1976, Torrance, 1987; Froberg and Kane, 1989; Kaplan *et al.*, 1993; Busschbach *et al.*, 1994; Bakker *et al.*, 1994; Silvertssen *et al.*, 1994; Gudex *et al.*, 1996). Gudex *et al.* (1996) report that of 3395 respondents only 3.2% of responses were excluded from analysis, whilst Silvertssen *et al.* (1994) and Ferraz *et al.* (1993) report completion rates of 95 and 100%, respectively. VAS methods have been found to be less expensive than other methods and quicker to complete, but not necessarily easier to complete (Torrance, 1976; Wolfson, 1982; Van der Donk *et al.*, 1995). VAS methods have performed well within clinical trials (Gabriel *et al.*, 1993; Bakker *et al.*, 1994; Busschbach *et al.*, 1994), across population studies (Gudex *et al.*, 1993; Gudex *et al.*, 1996) and also when presented via a multimedia format (Nease *et al.*, 1996). This high level of practicality has contributed to the wide use made of VAS methods.

Reliability

In an earlier review Froberg and Kane (1989) present evidence to demonstrate that VAS methods are reliable in terms of inter-rater reliability and test–retest reliability. Further to their review, more recent empirical results have reinforced the reliability of VAS methods (Gudex *et al.*, 1996; Bakker *et al.*, 1994; O'Brien and Viramontes, 1994; Gabriel *et al.*, 1994) (see *Table 1*). These studies cite test–retest correlation coefficients (intraclass correlation or Pearson correlation) ranging from $r = 0.61$ (O'Brien and Viramontes, 1994) to $r = 0.95$ (Bakker *et al.*, 1994).

Theoretical validity

There are supporters of VAS techniques who present them as cardinal measures of strength of preference (e.g. Kaplan *et al.*, 1993; Revicki, 1992), and indeed the techniques have been widely used in such a manner. The strongest theoretical argument for the VAS has been provided by Dyer and Sarin (1982), who suggest a measurable value function providing a link between such a value function and utility. Essentially the function represents preferences under certainty. Dyer and Sarin postulate that utility represents preferences under uncertainty and a stable relationship exists between values and utilities. Torrance and colleagues have interpreted this to provide a link between the VAS and SG, and hence a means of estimating SG values from VAS responses. This use of the VAS is examined later in this chapter.

The purpose of stated preference techniques is to imitate a real life situation to proxy RPs; however, VAS methods do not present a choice, and are therefore thought to be unable to measure strength of preference on a cardinal scale (Johannesson *et al.*, 1996). Due to the lack of choice and the absence of opportunity cost in the VAS task, one common view is that they have no basis in either economic or decision theory (Richardson, 1994; Nord, 1991).

The direct and choice-less nature of the VAS tasks has given rise to concerns over the presence of interval scale properties (Read *et al.*, 1984; Nord, 1991; Bleichrodt and Johannesson, 1997). There is a concern that VAS methods are susceptible to response spreading, whereby respondents use all areas on the valuation scale when responding (especially where multiple health states are valued on the same scale), which challenges VAS methods in terms of their interval scale properties. Response spreading (Parducci, 1974) can lead to health states which are very much alike being placed at some distance from one another on a valuation scale and health states which are essentially vastly different being placed very close to one another, as the respondent seeks to place (spread) responses across the whole (or a specific portion) of the available scale. If response spreading occurs, this provides another reason why VAS techniques may not generate an interval scale, and the numbers obtained may not be meaningful.

Empirical validity

Evidence: theory. The validity of the theoretical arguments presented by Dyer and Sarin (1982) have been considered, on empirical grounds, by Bleichrodt and Johannesson (1997). They carried

out a consistency test, examining whether valuations were independent of context, to see whether VAS methods offer a measurable value function. The authors found no support for the existence of such a value function. They report that VAS (RS) values were related to other health states included in the valuation task, where valuations were dependent on the numbers of health states that are preferred and less preferred to that state being valued. That is, the severity of the health states had no impact so long as the numbers of preferred or less preferred health states remained constant (similar findings were presented in a conference paper by Loomes *et al.* (1994)). Bleichrodt and Johannesson relate their findings to the possible impact of context effects and further support for findings of this nature is presented by Nord (1991) and Sutherland *et al.* (1983). If the interpretation of VAS valuations as points on a measurable value function is rejected, there remains no theoretical justification for the use of VAS method valuations in CUA.

Nord (1991) raises concerns surrounding the common practice of valuing numerous health states together on the same scale. Nord states that values may be affected by the states with which they are compared; he suggests that VAS scores should not be given much emphasis due to concerns over the equal interval scale properties of VAS methods. Torrance (1986) emphasises the need to stress to the subject that relative difference (between health states) is an issue, in order to capture interval scale properties when using VAS methods. Kaplan *et al.* (1993) have also recognised the potential for response spreading biases; however, they state that such biases may be controlled for by valuing one state at a time or through the use of a balanced design, and consequently the method may be not be flawed in this way.

Evidence: preferences. The ordinal properties of VAS methods have been largely accepted, given their history within psychometrics and psychophysics (McDowell and Newell, 1996; Gudex *et al.*, 1996). We have considered evidence concerning the convergent validity of the VAS and other techniques. VAS methods have generally been found to have only a weak correlation with SG and TTO (Rutten van Molken *et al.*, 1995; Clarke *et al.*, 1997; Zug *et al.*, 1995; Van der Donk *et al.*, 1995). There is insufficient evidence to infer any correlation between the VAS and ME or PTO. A further and possibly more important finding is that VAS methods correlate well with measures of health status (e.g. pain, functioning, clinical symptoms and instruments such as the SIP or AIMS – see chapter 6), to a much greater extent than SG or TTO. SG and TTO values have been shown to be only poorly to moderately correlated with measures of health status (Revicki and Kaplan, 1993; Revicki *et al.*, 1995). It would appear from the empirical findings of this review that VAS methods are measuring aspects of HRQoL which differ from those being considered by the SG and TTO.

In terms of hypothesised preferences, Gudex *et al.* (1996) report good VAS performance in terms of logical inconsistencies. Presenting results from the MVH study they report no logical inconsistencies in the rank order of median (or mean) valuations, and at an individual level found 57.4% of respondents had no logical inconsistencies at all. Gudex *et al.* using the 'strong' definition of logical consistency found the rate of inconsistency (with EQ-5D) to be 2.5%, based on the proportion of possible inconsistencies. These results would indicate a high degree of consistency in the VAS values obtained from a general population sample demonstrating a good performance against hypothesised preferences (i.e. based on EQ-5D descriptors). Further support is offered by an earlier population study (n = 287) undertaken by Gudex *et al.* (1993), which found an acceptable level of consistency with the VAS. In this study seven of 28 valuations (mean/median) showed a reversal of the logical orderings inherent in the Rosser disability and distress scale (controlling for factorial design). However, inconsistencies appeared to be related to one level of disability (V) which was described by a large amount of text, and this also caused problems for other techniques.

Given the evidence to suggest a strong correlation between the VAS and health status, it may be that VAS techniques are capturing more of the measurement aspect of health status changes than the satisfaction or benefit conveyed by such changes, whilst choice-based valuation techniques, such as SG and TTO, reflect the degree of satisfaction with movements in health status. This finding is supported by qualitative evidence of respondents seeing VAS methods as an expression of numbers in terms of 'percentages of the best imaginable state' (Nord, 1991), or a 'percentage of functioning scale' (Robinson *et al.*, 1997). If this is the case it would be understandable to see health status changes, which may not provide an overall improvement in HRQoL, reflected in VAS scores. On the contrary it may be that very small changes in health status result in a very large benefit, or value, to the individual, yet the actual benefit of such changes to the individual may not be reflected in VAS scores.

TTO review

Practicality

The TTO technique has proved to be a practical and acceptable method of health state valuation in a wide variety of empirical studies (Ashby *et al.*, 1994; Detsky *et al.*, 1986; Patrick *et al.*, 1994; Dolan *et al.*, 1996; Glasziou *et al.*, 1994; Johnson *et al.*, 1996; Fryback *et al.*, 1993; Kreibich *et al.*, 1996; Krumins *et al.*, 1988; Johnson *et al.*, 1996). For example, Johnson *et al.* (1996) report a 100% completion rate, Fryback *et al.* (1993) and Dolan *et al.* (1996a) report completion rates in excess of 95%, and Glasziou report a rate of 91% (see appendix 2 for further details). The TTO has been used in a self-administered form (Glasziou *et al.*, 1994; Perez *et al.*, 1997) with acceptable response rates, although most agree that interview application is preferable. Nease *et al.* (1995, 1996) have shown that computer-based applications of TTO are practical and acceptable.

Reliability

In their earlier review Froberg and Kane (1989) present evidence to support the reliability of the TTO technique. Further evidence of reliability has been presented by Dolan *et al.* (1996), Reed *et al.* (1993), Russell *et al.* (1992), Molzahn *et al.* (1995) Gabriel *et al.* (1994) and Ashby *et al.* (1994), with these studies presenting test–retest correlation coefficients ranging from $r = 0.63$ to $r = 0.85$ (see *Table 1*).

Theoretical validity

TTO has not been related in a specific way to any existing behavioural theory. However, the sacrifice element of the TTO task and its development from the SG afford it some common foundation in consumer theory (Dolan *et al.*, 1996; Johannesson *et al.*, 1994). Mehrez and Gafni (1990) restate the two-stage nature of the TTO, whereby the task first determines an indifference point between two certain periods of time (for the better and worse health states) and then secondly divides the shorter time period (logically representing the more attractive state) by the longer time period (reflecting the least attractive state) with the product of this division (i.e. x/t) representing a seemingly 'timeless' quality weight. In so far as the determination of the indifference point between two certain states goes, Mehrez and Gafni (1990) discuss the TTO in the context of value function theory, due to the identification of differing trade-off combinations of health and duration. Buckingham *et al.* (1996) align the TTO method with the welfare economic approach of 'compensating variation', where welfare gain is measured by compensating loss of something else that is valuable so that the respondent is returned to their original level of welfare. Although the theoretical arguments for the TTO are not as well developed as they are for SG, they emphasise the conceptual advantages of choice-based methods in economics.

The applicability of the TTO in medical decision-making has been questioned by some commentators (e.g. Mehrez and Gafni, 1991) due to the fact that the technique asks respondents to make a choice between two certain outcomes, when health care is characterised by uncertainty. Others (Stiggelbout *et al.*, 1994; Cher *et al.*, 1997) have argued that it is possible to adjust TTO values to incorporate individuals' attitudes to risk and uncertainty (see risk-adjusted QALY, chapter 2). Stiggelbout *et al.* (1994) present certainty equivalents* as a means of adjusting TTO values to account for uncertainty, whilst Cher *et al.* (1997) incorporate a parameter representing risk attitude to adjust TTO values to account for risk attitude with respect to gambles for survival duration. Richardson (1994) and Buckingham (1993) argue that the presence of uncertainty in the valuation task is not essential. Buckingham states that 'we do not need to express the value of risky prospects in terms of a risk' (p. 308), and Richardson points out that abstraction from risk *per se* is not a defect, the defect is abstracting from the risk of the health intervention itself.

An issue of theoretical concern in the TTO task is the effect of **duration**, that is, the effect of the specific time periods used in the task, hence relating to the circumstances of valuations. An underlying assumption of the TTO method is that individuals are prepared to trade off a constant proportion of their remaining life-years to improve their health status, irrespective of the number of years that remain. Yet the valuation of a health state may be influenced by the time an individual must spend it in that state (Sackett and Torrance, 1981). For example, individuals may adapt to health states and build up tolerance or they may become increasingly intolerant to that health state over time. Although duration can affect all techniques of health state valuation, given that the TTO requires a trade-off between two different time durations,

* Certainty equivalents ask respondents to identify the number of years in good health for certain they consider equivalent to a 50:50 gamble of a chance of a long or short length of life in good health. Certainty equivalents offer a measure of the utility for length of life and enable correction of TTO scores for effects of risk.

there are concerns over the effect of the specified duration on the underlying constant proportional trade-off assumption (Sutherland *et al.*, 1982). If individuals do not trade off a constant proportion of their remaining life expectancy in the valuation of health states, then values elicited using specific durations (e.g. 10 years) cannot be assumed to hold for states lasting for differing time periods, that is, they are not valued irrespective of the number of years that remain.

A further issue causing theoretical concern with the TTO technique is the impact of **time preference** on valuations, that is, **when** a health state occurs may be important to respondents (Dolan and Gudex, 1995). It may be that respondents would prefer to delay an episode of ill health or it may be that respondents would prefer to experience an episode of ill health immediately to get it out of the way. Such preferences can give rise to varying rates of time preference and contradict the assumption of constant proportional TTO. For example, if individuals have a positive rate of time preference they will give greater value to years of life in the near future than to those in the distant future. Where time preference has an impact on valuations it may not be valid to treat such valuations as an index of strength of preference for health. One solution to this has been a discounting procedure to incorporate time preference. By this method the value of each year is discounted by assuming a constant rate of time preference.

Empirical validity

Evidence: theory. Given the above discussion of the theoretical validity of TTO we have found empirical evidence to support the existence of time preference effects (Redelmeier and Heller, 1993; Dolan and Gudex, 1995) and to support the violation of the assumption of proportional TTO (Sackett and Torrance, 1981; Lipscomb, 1989; Stiggelbout *et al.*, 1995; Dolan, 1996). Dolan and Gudex (1995) present results which indicate wide variations in time preferences at the individual level, finding in a population sample (n = 39) that more responses implied negative rates of time preference than positive ones, although the modal time preference rate was zero. Furthermore, time preference rates are not constant, at least over wealth (Loomes and McKenzie, 1989). Considering the effect of duration, Sackett and Torrance (1981) found that health state valuations declined as the duration in the states increased (i.e. 3 months, 8 years and lifetime). Sutherland *et al.* (1982) found similar results with the VAS technique. Dolan (1996) in a population study (n = 234) found that the valuation given to a health state is a decreasing function of its severity and its duration. The author found evidence to support the hypothesis that 'dysfunctional health states will be seen as increasingly intolerable the longer they last'. Such evidence weakens the potentially pseudo-theoretical approach of the TTO method.

Evidence: preferences. A number of studies have considered the empirical validity of the TTO in a pragmatic fashion, on the basis of hypothesised preferences, that is, expected values (Dolan *et al.*, 1996b; Gage *et al.*, 1996; Churchill *et al.*, 1987) or stated preferences (Ashby *et al.*, 1996; Robinson *et al.*, 1997), finding evidence to offer some support in terms of empirical validity.

In terms of stated preferences, Ashby *et al.* (1996) compared TTO scores with the rank ordering of states given by respondents. They elicited health state valuations for health states of women after treatment for breast cancer, from a number of groups consisting of nurses, hospital doctors, general practitioners, university staff and breast cancer patients (total n = 138). Respondents ranked and valued five health states, presented with a health state baseline, and the authors report results that show a considerable degree of consistency in ranking and that rank ordering was consistently reflected in the mean TTO values.

Using convergent validity as an indication of TTO performance against stated preference we find there is evidence to suggest a reasonable correlation between TTO and SG valuations (Torrance, 1986; Dolan *et al.*, 1996b; Bosch *et al.*, 1996; Reed *et al.*, 1993; Zug *et al.*, 1995). However, whilst Torrance (1976) clearly believed TTO and SG to be equivalent, others (e.g. Read *et al.*, 1984; Wolfson *et al.*, 1982) have highlighted the differences between the scale values of the methods. Froberg and Kane (1989) remind us that good correlations do not necessarily produce equivalent scale values.

In a recent study Robinson *et al.* (1997) have considered differences between TTO and VAS responses by gathering qualitative data from 43 respondents who had taken part in a large scale health state valuation study. In this study it appeared that respondents had taken into consideration a wider range of issues in the TTO exercise than they did with the VAS. Sacrifice and duration were seen to be obvious issues for consideration in the TTO task. The respondents indicated that the TTO valuations were a better reflection of their health state preferences than were VAS scores, thereby reflecting on both the

theoretical and empirical validity of the TTO technique. Further qualitative evidence of this nature would be useful in the consideration of empirical validity.

Churchill *et al.* (1987) considered the quality of life in end-stage renal disease, using the TTO technique in a sample consisting of five patient groups in this area. These were hospital haemodialysis, home haemodialysis, self-care haemodialysis, continuous ambulatory peritoneal dialysis and transplant patients. The authors considered hypothesised preferences on the basis of the a priori prediction that transplant patients would score highest, hospital haemodialysis patients would score lowest and the others intermediate. The mean TTO scores for each treatment group confirmed these predictions, and the authors present these findings, in the absence of a reference measure, as circumstantial evidence of validity of the TTO.

Gage *et al.* (1996) elicited TTO valuations for three degrees of severity of stroke (i.e. mild, moderate and severe) in a sample of 70 patients. They assessed the validity of the TTO utilities indirectly by examining the ranking of the stroke utilities, expecting milder strokes to have higher utilities than more severe ones. The authors report that with only one exception the expected ordinal rankings of the stroke utilities occurred. Gage *et al.* (1996) also compared the TTO valuations for moderate stroke with those elicited using SG, and report that there was no significant difference between the utilities obtained between the two techniques.

As discussed under the SG review, Dolan *et al.* (1996b) assessed empirical validity of the TTO technique (and SG) on the basis of *a priori* expectations, developed from the results of previous studies. Results supported the hypothesis developed, with TTO and SG valuations reflecting expected preferences; however, readers are reminded of the contentious nature of the hypothesised preferences applied (see the SG review). Dolan *et al.* (1996b) also report that both TTO and SG (both with props and without) produced a similar ordinal ranking of health states.

Further evidence of the performance of TTO against hypothesised preferences is presented through reported performance in terms of consistency with logical orderings (i.e. expected preferences). Dolan *et al.* (1996b), in a general population sample of 335 respondents found TTO methods (both with and without props), to show high levels of consistency against 12 logically consistent comparisons (EQ-5D health state descriptors), finding a consistency rate of 91.7% for the original valuation task and again at retest. Gudex *et al.* (1993) in a population study (n = 287) found that six of 28 TTO valuations (mean/median) showed a reversal of the logical orderings inherent in the Rosser disability and distress scale used to elicit valuations. However, inconsistencies appeared to be related to one level of disability (*V*) which was described by a large amount of text. Dolan *et al.* (1996a), in a large general population sample of 3395, found the TTO method to be highly consistent. Laupacis *et al.* (1993) also present findings to indicate consistency in TTO methods.

One worrying aspect of the empirical findings of TTO studies is the extent to which respondents have been unwilling to trade or sacrifice any of their remaining life expectancy for improvements in health in some studies. In some situations such behaviour may be expected (this type of reaction is also seen in other choice-based techniques, see the SG review above), although in other instances it gives rise to concerns surround the willingness of the respondent to 'play the game' (Dolan *et al.*, 1996). Empirical studies have highlighted the potential 'unwillingness' of individuals to trade life expectancy in TTO tasks (Fryback *et al.*, 1993; Irvine *et al.*, 1995; Handler *et al.*, 1997; Robinson *et al.*, 1997). Irvine *et al.* (1995) report that 47% of respondents refused to trade-off any of their remaining life expectancy for a shorter life span in optimal health, despite experiencing a large number of health problems (ulcerative colitis and Crohn's disease patients, n = 94). Robinson *et al.* (1997) refer to a 'threshold of tolerability' below which health states would have to fall before respondents would be willing to sacrifice even a few days. Dolan *et al.* (1996) discuss the fact that studies can include those willing to trade quantity (life expectancy) for quality (health improvements), but find they do not wish to do so (eliciting a score of 1.00) and those unwilling to 'play the game' (also eliciting a score of 1.00), and they raise the question of how many of those health states scoring 1.00 may be due to preference and how many may be due to a refusal to participate. Dolan *et al.* refer to this as a qualitative difference between a TTO health state value of 1.00 and other health state valuations. In a population study, Dolan *et al.* (1996) found that approximately 5% of subjects were unwilling to sacrifice any life expectancy in order to avoid more than half of the states they valued. Furthermore, Dolan *et al.* found that respondents were more prepared to sacrifice life expectancy for states that included 'extreme

problems' with any of the dimensions of health presented. This latter finding may offer further support in terms of empirical validity.

PTO review

Practicality

The PTO technique (equivalence) has not been widely used to value health states. Applications of PTO have been in a methodological environment, and to date the feasibility and acceptability of PTO is relatively unknown. In an early examination of the technique, Patrick *et al.* (1973) report that it 'is too complex for use outside of a laboratory-like individual interview', stating that 'the task confused and offended some judges'. More recently, Nord (1993a) reports on two studies using PTO via self-administered questionnaires (using EQ-5D) in convenience samples of the Norwegian population and Australian students and nurses, with response rates of 28.2 and 27%, respectively. Nord (1993b) found that respondents in a convenience sample (n = 10) understood the PTO task (involving 14 pairwise choices), yet they found it difficult to choose precise equivalent numbers. Ubel *et al.* (1994) found that in 49 of 252 rationing choices, subjects thought that it would take an infinite number of people cured of the less severe condition to equal the benefit of treating ten in the more severe condition. Ubel *et al.* (1994) report that PTO was not easy to use (they administered PTO via a written survey). Nord (1995) reports that PTO can be quite demanding, warns of possible framing effects and advises the use of a multistep procedures to introduce individuals to the issues involved. Nord suggests that self-administered formats may not be suitable, and also advocates the use of a reflective element within PTO to allow individuals to consider their responses. Pinto Prades (1997) found PTO acceptable and feasible in a pilot study involving interviews with 30 undergraduate students. Murray and Lopez (1997) report the use of the PTO technique to elicit health state preferences for the assessment of the severity of disability as part of a large multinational study. Although Murray and Lopez do not present information on the performance of the PTO technique, they report that the PTO protocol was a group exercise (nine groups) for between eight and 12 participants lasting 10 hours, where discussion was an important aspect of the process and other health state valuation methods were used to encourage respondents to think carefully about the process.

As PTO asks individuals to consider choices concerning the treatment of others it is thought that subjects can find it difficult and unpleasant to make such direct decisions (Nord, 1995). In a pilot study involving 53 Norwegian politicians Nord (1995) found that there was a willingness to respond to PTO tasks (36 respondents); however, Nord also found evidence that there was some reluctance to participate (17 respondents). Further evidence is required from empirical studies to demonstrate the practicality of PTO.

Reliability

As with practicality, the reliability of PTO is yet to be demonstrated. Patrick *et al.* (1973) report a comparatively low correlation (Pearson; $r = 0.60$) with respect to inter-rater reliability, compared with the VAS (RS) and ME ($r = 0.75$–0.77 and $r = 0.75$–0.79, respectively). Further to this, Nord (1993, 1995) finds a strong random element in individual PTO responses, but suggests responses may be reliable at a group level. Nord (1993b) reports retest findings from 20 individual PTO responses ('some weeks after the first response') showed a mean difference of 40%. However, there is not much evidence on reliability, and Nord himself advocates further research in this area.

Theoretical validity

There presently appears to be no formal theoretical support, within economic or decision theory, to underpin the PTO technique. Although the technique is seen as intuitively appealing (Nord, 1995) there are no theoretical underpinnings advocated in the current literature, other than psychometric qualities surrounding adjustment or equivalent stimuli (Patrick *et al.*, 1973). Patrick *et al.* (1973) talk of PTO as a technique whereby subjects express preferences in terms of a point of subjective equality or indifference. Although the PTO technique is choice based, the choice is made in a 'social' context with outcomes relating to the welfare of others; therefore, standard consumer theory cannot be applied to the decision task. It may be that PTO, due to its social preference perspective, can be linked to the economics literature surrounding the valuation of externalities, yet this has not been the case so far.

Richardson (1994) supports the potential interval scale properties of PTO due to the fact that there is a clear and comprehensible meaning to PTO (where the numbers are specified). Pinto Prades (1997) also comments that one of the hypothetical advantages of PTO is that it asks the right question (i.e. trade-offs between people).

Empirical validity

Evidence: theory. As a measure of social preference, PTO has been used to validate other

techniques of health state valuation, in the context of resource allocation decisions (Nord, 1991; Nord *et al.*, 1993). However, in terms of assessing the performance of the PTO itself few studies have been undertaken.

In considering the use of a reference outcome such as the SAVE, Ubel *et al.* (1994) have presented some preliminary findings, from a self-administered survey, to suggest that PTO were not internally consistent. They report that the results of a test for multiplicative transitivity, whereby responses against a standard unit such as the SAVE (they use treatment for acutely fatal appendicitis) should have a consistent relative value, show PTO responses as a whole were not consistent. The authors find that 'in the majority of cases the indifference points predicted by multiplicative transitivity were greater than the values the subjects gave when given direct rationing choices'. Although the authors do highlight some reasons to be cautious about the results they report, they also warn that in order to infer relative values for PTO pairwise comparisons the technique must be internally consistent.

Evidence: preferences. The study by Ubel *et al.* offers some insight into the performance of the PTO technique against stated preferences. Further empirical evidence of a stated preference nature can be found in the convergent validity of the PTO with other valuation techniques. Patrick *et al.* (1973) report no significant differences between valuations elicited using PTO (equivalence), ME and the VAS. Froberg and Kane (1989) cite a study by Miles (1977) – not obtained for inclusion in this review – which compared the PTO with other scaling methods, reporting the VAS (CR) and PTO (equivalence) as having nonsignificant differences across 12 comparisons.

With respect to hypothesised preferences, Pinto Prades (1997) compared the performance of PTO, SG and the VAS on the basis of predictive power, measuring the degree of ordinal agreement between an expected ordering (given by the respondents) and an ordering directly obtained using the three methods at an individual level (the degree of agreement was measured using Kendall's measure of association – where 0 reflects independence between orderings). The VAS was found to have a poor association (–0.06), whilst SG and two of the three variants of PTO used were found to have a better association, having similar results (0.34–0.393), One of the PTO variants (PTO-3) was found to have a greater level of association (0.621) than the other methods used. Pinto Prades (1997) also assessed the techniques on the basis of strength of preference (cardinal) using a hypothetical voting exercise reflecting the treatment intervals of paired comparisons, finding that the PTO (variant 3 used) was a better reflection of social preferences than other techniques used. The study by Pinto Prades is a pilot study within a convenience sample, yet it may suggest that PTO is better able to reflect social preferences than are SG and VAS methods. However, some would argue that these latter methods do not set out to measure social preferences.

When considering consistency, as a measure of performance against hypothesised preferences, again we have found that there is limited evidence covering PTO responses. Ubel *et al.* (1994) report that 11 of 53 respondents' PTO scores were excluded from analysis due to a large number of inconsistent responses, and of the remaining 42 respondents, inconsistent responses were common. Some of the inconsistency experienced by Ubel *et al.* may be due to responses being elicited via a self-administered survey.

Nord (1992) comments that when the PTO technique is used, distributive considerations can become a serious confounding factor, and these may limit weights between serious and less serious conditions. Nevertheless, Nord (1993a, 1994) hypothesises that social values elicited via PTO responses will reflect a societal preference for life-saving interventions relative to health-improving interventions. Ubel *et al.* (1994) and Pinto Prades (1997) report results which support such a hypothesis, finding that PTO responses consistently reflected a respondent preference to treat those people whom they saw in a worse state. Nord and Ubel (1994) talk of the 'rule of rescue' (Hadorn, 1991), reflecting a desire to save identifiable people from significant distress.

The literature reviewed here has not considered the effect on PTO responses of general issues, such as attitude to risk and the effect of duration or time preference. Nord (1995) has stressed the importance of considering possible biases, and has suggested that 'PTO response may be sensitive to the arguments mentioned in the questions, to the choice of start-point, the numbers in pairwise comparisons, and to the choice of decision context'. As highlighted earlier, PTO responses may be a function of many considerations, for example initial severity, health status after intervention or the nature of the intervention, and at present there is no empirical evidence to disentangle the responses, to determine the important arguments and their relative weights. Nord (1995)

recommends a 'reflective equilibrium' approach, which takes respondents through a multistep procedure, in order to carefully consider the relevant arguments and to reconsider initial responses in the light of their implications.

ME review

Practicality

Froberg and Kane (1989) state that along with the VAS, 'magnitude estimation is the least expensive and easiest [technique] to understand', and they also indicate that ME has provided high response rates, yet they fail to quantify or reference such statements. Further to this, we have found no reports of either response rates or completion rates for ME in any of the literature identified in the search.

Reliability

Patrick *et al.* (1973) in commenting on internal consistency reported intrarater coefficients (Pearson correlation) of 0.74–0.83 for ME, whilst inter-rater reliability was given as 0.75–0.79. Rosser and Kind (1978) reported test–retest reliability (as measured by percentage agreement) at 97.2% whilst inter-rater reliability was 88%. Gudex *et al.* (1993) found that of the three instruments used in their study (ME, the VAS and TTO) ME was the instrument most affected by interviewer bias (indicating poorer inter-rater reliability in ME than in the VAS and TTO).

Theoretical validity

ME is not a choice-based task and therefore consumer theory can offer no theoretical support. ME has not been related to any behavioural theory, and we have found no theoretical underpinnings for the technique in the economics literature. Stevens (1971) championed ME as a way of overcoming the weaknesses of the VAS (CR); in particular, ME was put forward as a way of tackling the lack of ratio level measurement in VAS (CR) techniques. The key assumption regarding ME as an instrument for producing health state valuations is that it produces a ratio scale (although it is generally accepted that interval-scale properties are sufficient for health state valuation instruments). Patrick *et al.* (1973) support the claim that ME is a ratio measure; however, there are some serious doubts expressed as to whether this is a tenable assumption. Kaplan *et al.* (1979) are not convinced of the ratio scale properties, and Richardson (1994) is also sceptical of the ratio properties of ME. He states that the meaning of the ME question, of how many times is *x* worse (better) than *y*, is 'deeply obscure'.

The choice of anchor (i.e. the choice of the health state defined as 0), has posed some problems in ME applications. Kaplan *et al.* (1979) used death as equal to 0, whilst in their study Haig *et al.* (1986) inverted the 0 to 1 scale by using the absence of dysfunction and discomfort as 0. In this instance discomfort is defined as pain/anxiety, combined with duration and intensity.

Empirical validity

Evidence: theory. Assessing the assumption of ratio scale properties, Kaplan *et al.* (1979) report empirical findings to suggest that ME does not have ratio properties. They report the results from a sample of 65 college students rating 30 health state descriptions, in which ME was found to give values which were compressed to the lower end of the measurement scale (death). On this basis (as well as other evidence comparing ME with the VAS) Kaplan *et al.* assert that ME is inappropriate as a measurement method for a health state index. Findings from Haig *et al.* (1986) are more supportive of the ratio properties of ME. They report that the study by Kaplan *et al.* (1979) introduced 'floor effects' through the use of death as 0 (in effect not allowing states worse than death to be valued), and they instead use the absence of dysfunction and discomfort as zero (as mentioned above) to attempt to overcome such effects. Haig *et al.* report that by using their method ME does indeed have ratio properties. Rosser and Kind (1978) were also convinced that ME has the properties of a ratio scale and used ME to value the 29 states from the Rosser and Kind matrix. Although they added the caveat that more work was needed to verify such a claim.

Evidence: preferences. With respect to empirical evidence reporting findings against stated preferences, Patrick *et al.* (1973) found no significant differences in the values elicited via ME, the VAS and PTO (equivalence), presenting this as an indication of the convergent validity of the three methods. The authors also stated the relation between the VAS and ME to be 'clearly linear'. Gudex *et al.* (1993) report findings from a population study (n = 287), which suggest convergence between ME and the VAS, although they also present information indicating differences between the techniques.

In the study by Patrick *et al.* (1973) there are some doubts cast on their findings in this instance. They restricted the scale at the upper end with the description of a 'perfectly well day'. This is not the usual practice in ME, and in doing so they may have effectively turned ME into a VAS task (scale of 0 to

100). In this case it is hardly surprising that some form of ME and VAS convergence was shown. Kaplan *et al.* (1979) also cast doubt on the convergent validity of ME with the VAS (CR). In their study the compression of scores at the lower end of the scale meant that these findings were inconsistent with the VAS (CR). Although as already pointed out above, the 'floor effects' that Kaplan *et al.* introduced by using death as 0 may explain these inconsistencies.

Gudex *et al.* (1993) offer some evidence of ME performance against hypothesised preferences. They report that ME produced mean/median values which were generally consistent with the logical orderings inherent in the Rosser disability and distress scale, finding only three of 28 mean/median valuations with a reversal of the expected orderings. The only further evidence found to support the performance of ME against this form of empirical validity comes from Kind *et al.* (1982) – cited by Froberg and Kane (1989) – who found that the value of health states as implied by the relative values of UK court awards for personal injury claims was significantly correlated with ME scale values.

We found the empirical literature covering ME to be very limited and failed to identify any examples of ME being used primarily in a clinical study. All the articles found were concerned either with comparing ME with various other measures, investigating the fundamentals of ME, or in using ME to provide tariff values in the case of the Rosser and Kind matrix.

Comparison of health state valuation techniques

Practicality

The five methods of health state valuation discussed in this review have generally been reported to be practical and acceptable (Froberg and Kane, 1989). This in part reflects the development of props, training of interviewers and other aspects of good quality in the administration of the techniques (Torrance, 1986). We have found a lack of empirical evidence to demonstrate the acceptability of ME and PTO. The lack of evidence for ME reflects the limited use made of ME in health care. We have confined our search to the area of health and health care, and it may be that more extensive search strategies can further inform on the application of ME. The lack of evidence concerning the PTO reflects the relatively short history of using the technique.

We report empirical evidence to support the acceptability of SG, TTO and VAS methods, although there is some variance in the findings. VAS methods tend to outperform both SG and TTO. In a number of studies using the VAS and SG and/or TTO, VAS methods have had a greater level of completion (e.g. Detsky *et al.*, 1986; Van der Donk *et al.*, 1995; Revicki 1992; Patrick *et al.*, 1973).

As one would expect, given the more complex cognitive task, it would appear that the choice based SG and TTO techniques result in a larger number of refusals, missing values and inconsistent responses. Some of these completion difficulties are due to misunderstandings, conflicts with personal beliefs or straightforward difficulties in understanding the tasks (Fryback *et al.*, 1993, Gage *et al.*, 1996). Other studies have reported that respondents have difficulties dealing with small probabilities (Cairns *et al.*, 1996) or overestimate small probabilities (Loomes and McKenzie, 1989). As discussed in the review of SG and TTO, although respondents find the techniques acceptable they are often unwilling to make a sacrifice in the valuation task. That is, respondents have been unwilling to trade any of their remaining life expectancy (not even a few days or hours) or have been unwilling to accept any level of risk (not even a very small level of risk), to obtain an improved health state, even when respondents are experiencing a large number of health problems (e.g. Handler *et al.*, 1997).

Although SG and TTO have demonstrated a similar performance in terms of completion, TTO has outperformed SG by a small margin in a number of studies (e.g. Reed *et al.*, 1993; Van der Donk *et al.*, 1995; Dolan *et al.*, 1996).

Earlier reviews (Froberg and Kane, 1989) have reported that SG is complex and not intuitively obvious to most respondents, stressing that it may be too complex for population studies. We have reported empirical evidence to support its successful use in population studies (Dolan *et al.*, 1996b, n = 335). The TTO technique has also proved to be a practical technique in population studies (Dolan *et al.*, 1996a,b). Both SG and TTO have shown to be feasible in a self-completed format (TTO, Glasziou *et al.*, 1994; SG no props, Dolan *et al.*, 1996b), but both techniques are unproven in postal format. The VAS technique has been used more widely with broader evidence of acceptability (e.g. Essink-Bot *et al.*, 1990; Silvertssen *et al.*, 1994).

There has been a growth in the number of studies utilising computer-based methods for the administration of valuation tasks. We have reported empirical evidence (see appendix 2) to show

that computer-based methods (e.g. the 'U' titre, Nease *et al.*, 1995, 1996) have proved feasible and acceptable for the presentation of SG, TTO and VAS tasks (Krahn *et al.*, 1994; Morss *et al.*, 1994; Nease *et al.*, 1995, 1996; Clarke *et al.*, 1997).

SG and TTO have been found to be practical on most populations, but the VAS is usually marginally better in terms of response rate and cost.

Reliability

Froberg and Kane (1989) report an acceptable level of intrarater reliability for all five of the techniques discussed and good to moderate levels of inter-rater reliability for VAS (RS), ME and PTO (equivalence). They do, however, comment on the general lack of evidence surrounding the reliability of methods. Ten years on this problem still exists, with many studies either failing to undertake or failing to report tests of intrarater and inter-rater reliability. Given the data available *poste* Froberg and Kane, we have focused on reliability over time, test–retest reliability, and presented results in *Table 1* and appendix 2.

There is a lack of evidence surrounding the test–retest reliability of PTO and ME techniques. However, of the two techniques ME would appear to be the most promising in terms of reliability (Rosser and Kind, 1978). Rosser and Kind report test–retest reliability for ME at 97.2%, measured by percentage of agreement. Nord (1993b) report poor test–retest findings for the PTO (40% measured by percentage of agreement), and express concerns over a strong random element in individual PTO responses. Further developments continue with the PTO technique and further empirical evidence is emerging.

Table 1 reports empirical evidence covering SG, TTO and VAS techniques, with all of these demonstrating an acceptable level of reliability. In three of the five comparative studies reported in the table, however, choice-based methods outperform the VAS, with one case showing similar results at 1 week retest. SG and TTO techniques display similar results across comparative studies in terms of reliability. Empirical evidence would suggest that although it would be difficult to express a preference over the two techniques on the basis of reliability, the TTO offers slightly better performance statistics on test–retest reliability, as can be seen by its greater reliability in three of the five comparative studies cited in *Table 1*. Yet, as shown in the table, the differences between the reliability of all three techniques are small.

There is little to choose between the SG, TTO and VAS techniques on the grounds of reliability.

Theoretical validity

We have critically reviewed the basis of each technique in economic theory and conclude the following.

Although it has been argued that VAS techniques are a means of eliciting a measurable value function (Dyer and Sarin, 1982), such a theoretical basis is not established in economics. The VAS does not present the respondent with a choice, and has no element of sacrifice or opportunity cost, thereby leaving economists with no means of applying consumer theory and decision theorists with no means of predicting decisions. This leaves the foundation in psychometrics and psychophysics, which have no direct link with the measurement of strength of preferences.

Although ME has a fairly long history in the healthcare decision-making literature (Stevens, 1971) it has been largely unused and theoretically undeveloped. ME does not present the respondent with a choice, and some find its meaning obscure and inappropriate (Richardson, 1994). The theoretical appeal of ME rests on an assumption that it is able to provide ratio scale properties (Stevens, 1971; Patrick, 1973; Rosser and Kind, 1978), yet there are serious doubts concerning the basis of this assumption (Kaplan *et al.*, 1979). ME is not related to a behavioural theory, and the assumption surrounding its ratio level properties remains unsupported.

TTO is a choice-based technique, involving sacrifice and opportunity cost, and as such may find some association with consumer theory and welfarism. Theoretical support has been sought amongst those theories surrounding equivalent and compensating variation; however, such support is not developed within the current literature, and TTO remains unrelated in a specific way to any behavioural theory. Should TTO be considered in the context of consumer theory three concerns would present themselves. These being the effect of duration, the impact of time preference and the incorporation of uncertainty.

PTO is a choice-based technique, but it relates to social choice, and the opportunity cost is not directly borne by the individual. Therefore, consumer theory is not applicable in this instance. PTO is intuitively appealing, and its approach is considered by some to be meaningful (Richardson, 1994). It may be that theoretical underpinnings

will be developed within the economics literature, with potential opportunities in the externalities and social choice literature.

SG is undoubtedly the most theoretically appealing of the techniques reviewed here. It has rigorous theoretical foundations in the form of EUT, which has proved to be the dominant theory of decision-making under uncertainty since the 1950s. Although its restrictive axiomatic approach has many critics (e.g. Loomes and McKenzie, 1989), EUT has demonstrated its theoretical application to choice under uncertainty, and due to the uncertain nature of medical decision-making finds support for its application to health state valuation (Torrance, 1986; Gafni, 1994). Due to its link with EUT, SG has been put forward as the criterion or reference method of health state valuation, often referred to as the 'gold standard'. However, in the face of theoretical arguments against the 'favoured' use of EUT and SG in health state valuation and in consideration of the limited empirical support for the application of EUT (discussed below) we cannot support the 'gold standard' status of the SG.

From a theoretical viewpoint only, choice-based techniques should be used: SG, TTO and PTO. The choice between SG, TTO and PTO depends on the perspective employed. The debate surrounding SG versus TTO is unresolved and depends on ones belief in the descriptive validity of the theory or its normative basis.

Empirical validity

Evidence: theory. We have reported that both PTO and ME lack theoretical support, and have therefore been unable to comment further on these techniques. The theoretical argument associated with the VAS technique, that is, measurable value function, has been challenged by Bleichrodt and Johannesson (1997), who present findings to suggest such a function is not present. We have also reported evidence to suggest that the presence of response spreading in VAS methods is a significant concern. Although TTO is not directly linked to specific theoretical foundations, we have reported evidence to suggest that duration effects and time preference effects can have an impact on the elicitation and use of TTO values (Sutherland *et al.*, 1982; Dolan and Gudex, 1995). We report that it is possible to adjust TTO-elicited values to address the absence of uncertainty (Stiggelbout *et al.*, 1994; Cher *et al.*, 1997), yet we have found no further empirical literature to demonstrate such adjustment.

SG is the only technique with clear theoretical foundations (i.e. EUT), but we have reported evidence showing that SG values can be strongly influenced by the outcomes used in the task (i.e. non-independence) and by the manner in which the task is presented (Llewellyn-Thomas, 1982; Gage *et al.*, 1996) and we have reported evidence to suggest that attitude to risk is not constant (e.g. Kahneman and Tversky, 1982; Loomes and McKenzie, 1989; Stiggelbout, 1995; Clarke *et al.*, 1997). Such evidence suggests that the axioms of EUT are often violated. If the axioms of EUT are empirically flawed (as it relates to health state valuation) as many commentators suggest, there can be no justification for SG as the reference method, or 'gold standard', for health state valuation.

Evidence: preferences. We have not found a large literature reporting the empirical validity of valuation techniques. Researchers have for some time found the assessment of techniques in terms of empirical validity a difficult task, this being due to the absence of any reference unit of measurement, such as RPs. Yet, we have been able to report evidence which assess empirical validity of techniques on the basis of their performance against stated and hypothesised preferences.

In terms of stated preferences we have considered techniques in relation to available measures of stated preference, for example ordinal ranking of health states, and also in the context of valuations elicited from other techniques employed at the same time (i.e. convergent validity). Although many studies consider such issues in their protocol, we have found that often such results are not reported (e.g. Dolan *et al.* (1996a,b) and Gudex *et al.* (1996) do not report the results of ranking tasks which may offer another form of stated preferences).

There is evidence of a poor to moderate correlation between VAS values and values from choice-based techniques undertaken at the same time (i.e. SG and TTO). This finding, together with significant evidence to suggest a strong correlation between the VAS and measures of health status, raises concerns over the ability of VAS methods to elicit strength of preference for health states (these concerns are also compounded by the findings relating to response spreading). Such concerns are further supported by qualitative data reported in the study undertaken by Robinson *et al.* (1997), where respondents indicated that their VAS responses did not truly reflect their preference. Although the ordinal properties

of VAS methods are largely unchallenged, findings of this nature cast doubt on whether VAS values are able to reflect respondent strength of preference.

There is little evidence concerning the performance of ME and PTO against stated preferences. Gudex *et al.* (1993) report a comparison between ME, the VAS and TTO, with some evidence of convergence between ME and the VAS, although overall they cite tests (Friedman test statistic) indicating important differences between valuations produced by the different methods. These studies are limited, and further evidence is required to support any relationship.

SG and TTO values have been found to correlate reasonably well with one another (e.g. Torrance, 1986; Dolan *et al.*, 1996b), suggesting they may be valuing similar aspects of HRQoL. Whilst we have found no direct evidence to inform on the performance of SG against stated ordered preferences, Ashby *et al.* (1994) report that TTO produced mean values which were consistent with respondent rank ordering of health states and Robinson *et al.* (1997) present qualitative evidence to indicate TTO responses reflect the stated preferences of individuals.

We report evidence relating to consistency of response with a priori preferences, and find that SG, TTO and the VAS have demonstrated good levels of consistency with the MAUS, whilst a lack of evidence surrounding PTO and ME leaves the consistency of these methods unproven. Although consistency may be a function of the medium used to present health states for valuation (e.g. the EQ-5D), thereby making the comparison of methods dubious, two of the studies reviewed report comparative results. Gudex *et al.* (1993) report that consistency of TTO responses was superior to the VAS, although the difference was small (reversals of logical ordering found in six of 28 TTO responses as opposed to seven of 28 VAS responses), and ME was found to be superior to both of these techniques (with only three inconsistent responses of 28). Gudex *et al.* do however report that inconsistency may have been related to one particular level of disability. Dolan *et al.* (1996b) report a TTO consistency rate (91.7% for props and no props versions) superior to SG (83.8–87.5%). Gudex *et al.* (1996) and Dolan *et al.* (1996a) report high levels of consistency, from the VAS and TTO, respectively, amongst a large general population sample (n = 3395); however, quantitative information is not presented in the latter study to allow a direct comparison.

Further to the above study by Gudex *et al.* (1993) the consistency of ME is not discussed in the literature reviewed. The consistency of PTO responses is also relatively untouched; however, Ubel *et al.* (1994) have reported (self-administered survey) a high number of inconsistent responses and concerns over the internal inconsistency of PTO, that is, in relation to a relative unit of measure.

A number of studies have reported the performance of SG and TTO against hypothesised or expected (*a priori*) preferences. The study by Gage *et al.* (1996) reports that TTO values reflected the expected ordinal ranking of stroke severity (mild, moderate, severe) and that TTO and SG values for moderate stroke were not significantly different (SG used to value moderate stroke only). Dolan *et al.* (1996b) have reported that both SG and TTO performed well against hypothesised preferences with respect to the background characteristics of respondents. Churchill *et al.* (1987) have reported that TTO values reflected hypothesised preferences within the valuation of end-stage renal failure treatment modalities.

There is very limited evidence on empirical validity. In relation to consistency with the MAUS, the evidence marginally favours TTO, but this is not sufficient to say one technique is more valid than another on empirical grounds.

Summary of previous reviews

As we have shown, a number of commentators have previously compared some or all of the health state valuation techniques discussed in this review. Whilst some authors have drawn attention to the fact that the comparison of techniques is difficult given the number of variants of each technique (e.g. Nord, 1992; Dolan and Sutton, 1997) and given study-specific considerations (e.g. Nord, 1992; Ferguson and Keowen, 1995), others have stated their preferred techniques. We offer only brief details of such preferences, and advise the reader to consult the original references for further guidance.

Torrance (1976) compared SG, TTO and the VAS, and points to 'the TTO method as the best of the three tested for use on the general public', with SG 'as a close second'. Torrance, commenting on the VAS (CR), found that 'its only redeeming virtue is its potential lower cost'. Torrance reaffirms this view in his 1987 paper by stating a preference for TTO over SG, recommending that researchers should use TTO if they can afford it and the VAS (RS) with a power curve correction if they cannot. However, Torrance has more recently come out in

favour of the SG technique (Feeney and Torrance, 1989). Wolfson *et al.* (1982) considered SG, TTO and the VAS and favoured TTO, as it would 'seem more promising', given problems with the VAS and difficulties in the application of SG. Froberg and Kane (1989) reviewed a wide range of techniques (including those reviewed in this report) and concluded that 'based on data concerning reliability, validity and feasibility, the most promising scaling methods are the category ratings [VAS], magnitude estimation, and the time trade-off methods'. Mehrez and Gafni (1991) find that 'because medical interventions occur only in a world of uncertainty the SG is the appropriate technique, they reaffirm this belief in 1993. Richardson (1994) reviews the techniques from a welfarist perspective, and finds the TTO and PTO more satisfactory than the SG, VAS and ME techniques. Dolan *et al.* (1996b) considered SG and TTO in both props and no props form, and found the TTO props versions to be better than TTO no props and TTO to slightly outperform SG (TTO better on completeness, marginally better on logical consistency, significantly better on reliability), but state that there was no clear-cut 'winner' from their study and that there was little to choose between the TTO props and SG props techniques.

Mapping from the VAS to SG or TTO valuations

Introduction

The VAS technique may not produce health state utilities for calculating QALYs, but there has been interest in mapping VAS values to SG or TTO utility values. Torrance *et al.* (1992), for example, used the VAS to elicit preferences for the single and multiattribute health states defined by the HUIs and transformed these values into SG utilities using a specially estimated power function (see chapter 3). The advantages of the VAS are that it presents a more familiar task and respondents report finding it easier to complete than the choice-based methods. The review presented earlier found evidence of better rates of completion, consistency and reliability being achieved by the VAS than SG or TTO. The VAS also presents fewer ethical problems than SG or TTO, particularly with patient groups, since it does not present respondents with potentially upsetting scenarios involving death (Drummond and Davies, 1991). There would be significant practical advantages to being able to map from the VAS to one of the choice-based techniques.

Torrance (1976) first observed that TTO and SG values exceeded VAS values and claimed the relationships were curvilinear. Represented on a plot with SG or TTO values on the vertical and VAS values on the horizontal axes, points would be above a 45° line and bow outwards, as shown in *Figure 5*. Torrance estimated a power relationship between the VAS and TTO for health state mean values, and this has been replicated by himself between the VAS and SG (Torrance *et al.*, 1996) and by other researchers (e.g. Loomes, 1993; Stigglebout *et al.*, 1996; Bleichrodt and Johannesson, 1996; Dolan and Sutton, 1997).

The question addressed in this section is whether the relationships are sufficiently robust to use the VAS to accurately and reliably predict SG or TTO values. This section reviews the theoretical grounds proposed in the literature for any relationship between the VAS and these two choice based techniques. This is followed by a review of the evidence for unique, statistically sound and stable relationships between the VAS and SG and the VAS and TTO.

Theoretical explanations

There are numerous theoretical explanations for the relationship between the VAS and the choice based methods; some are unique to SG or TTO and others are common. This review describes them in order to help interpret the evidence presented later and to consider the likely stability of any estimated relationship.

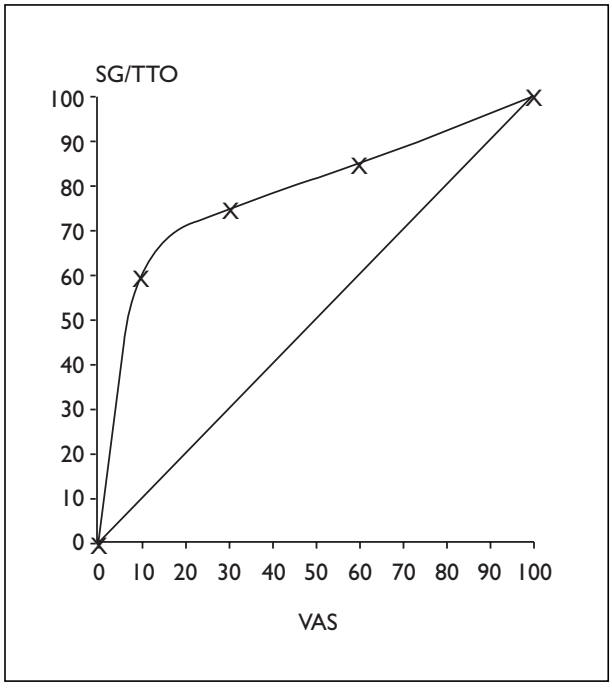


FIGURE 5 *Predicted relationship between SG and VAS*

Risk attitude

The main theoretical argument for the relationship between the VAS and SG has been from the work of Dyer and Sarin (1982), since used by Torrance and colleagues in their work with the HUI-II and HUI-III, who argue that utilities are a combination of the measurable value function and relative risk attitude (see earlier). They regarded the VAS as a technique for eliciting this measurable value function and SG the technique for eliciting utilities. According to this explanation, SG and the VAS will only be the same for individuals who are risk-neutral. A risk-averse person would exhibit a concave relationship between the VAS and SG, 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). For the risk seeker, it would have a convex shape indicating the opposite. Given most people have been found to be risk averse over uncertain prospects involving health this provides an explanation for the relationship presented in *Figure 5*.

Torrance has proposed the following power function to represent this relationship:

$$U = 1 - (1 - V)^{b} \qquad (1)$$

where the power term b is a person's constant relative risk attitude with $b > 1$ implying risk seeking, $b < 1$ implying risk aversion and $b = 1$ risk neutrality, U is the SG 'utility' and V is the VAS value.* A crucial feature of this function is that risk attitude is assumed to be constant.

Bombardier *et al.* (1982) explained the relationship between the VAS and SG 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'. This general aversion is fixed regardless of the level of risk, and hence is represented by a constant term for the difference between SG and the VAS. Combined with relative risk aversion it implies the constant term in equation (1) may not equal unity. The existence of a fixed gambling effect in SG utilities has been acknowledged by a number of health economists (Gafni, 1994; Richardson, 1994). Richardson (1994) has argued that this gambling effect is not allowed for in EUT, and offers a different or complementary explanation to relative risk aversion.

Time preference

There has been rather less discussion in the literature on the expected relationship between the VAS and TTO. An obvious source of any difference is time preference. A health state value is obtained from TTO by dividing the length of time in the chronic state (t) by the selected lesser time in full health (x). The time difference has no implications for a zero rate of time preference (analogous to risk neutrality for SG). However, any other rate of time preference would 'contaminate' the result. A constant positive time preference rate, as assumed in financial investment analysis, would reduce the value of time spent in the chronic state (i.e. time t) by a greater proportion than the time spent in full health (i.e. x), and hence the ratio of x to t would increase. For a constant positive rate of time preference, the size of the difference between discounted and undiscounted TTO health state values is maximised over the middle range and hence produces the characteristic outward bowing shown in *Figure 5* (see appendix 2 for further detail). A negative time preference rate would have the opposite effect.

Duration

It has been suggested that time may have another consequence for TTO health state values through the impact of duration. There is evidence that for the prospect of poor health states, particularly severe ones, they seem worse the longer they are specified to last in the valuation task (Torrance and Sackett, 1978; Sutherland, 1982). This would result in TTO values declining with time, which would have the opposite effect of a positive time preference. The extra time in the chronic state (i.e. time t) would weigh more heavily than the time in full health and hence result in a lower TTO value. Conversely people may believe that with time they will adjust to the state and hence the duration effect may raise the TTO health state values. Existing evidence on valuing hypothetical states supports the former hypothesis.

Framing effect

An important explanation for the difference between the VAS and TTO or SG has been the use of different reference points in the valuation tasks (Loomes *et al.*, 1994; Dolan and Sutton, 1997). In the SG question, a respondent is asked to imagine that he/she is in a chronic health state, certain to last for some period of time. He/she is then asked to consider a risky treatment option, one of which

* Currim and Sarin (1984) have proposed another possible relationship: $U(X) = (1 - e^{-cv})/(1 - e^{-c})$. The parameter c is a constant, and reflects the person's relative attitude to risk as follows: they are relatively risk averse when $c < 0$; risk neutral when $c = 0$ and risk seeking when $c < 0$.

is better than the reference state and the other worse. The chronic state of each SG question therefore becomes the reference state. The TTO question also asks the respondent to imagine they are in some chronic state. Whereas with the 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).

According to Kahneman and Tversky's (1979) prospect theory, gains or losses relative to their perceived reference point are valued differently. Kahneman and Tversky have argued that individuals are risk averse over gains and risk seeking over losses. Under the VAS, any departure from full health would be regarded as a loss, and hence would have a lower value than either SG or TTO. Loomes *et al.* (1994) have shown how Kahneman and Tversky's value function, combined with the different reference points of the VAS and SG, can generate a curvilinear function between SG and the VAS. However they go on to show that unlike the relative risk attitude explanation, it does not imply a constant power term. This 'reference point plus value function' would suggest that a separate power function would have to be estimated for each health state.

Overview

There is no single explanation for the relationship between the VAS and SG or between the VAS and TTO. Some theories suggest a potentially stable functional form for the relationships. However, the theories suggest different forms and in one case a different direction for the relationship. For the VAS to SG relative risk attitude implies a power function for the relationship, while the gambling effect suggests a linear function. For the VAS to TTO a constant time preference rate results in a power function, but the impact of duration could be in the opposite direction. It is likely these theories will operate simultaneously, and this presents estimation problems.

These theories also depend on individuals' preferences conforming to some assumption, such as constant relative risk aversion. One attempt to introduce a more complex theoretical explanation for the differences is the application of reference point theory and this suggests a need to estimate a separate function for each health state being valued. This result casts some doubt on the chances of estimating a single relationship at all.

These explanations also presuppose that the VAS can generate a measurable valuation function for strength of preference over health states under certainty and SG can measure the utility value of states under uncertainty. Serious doubts have been raised about the ability of the VAS to estimate a value function (Bleichrodt and Johannesson, 1997; Loomes *et al.*, 1994). This would suggest that any relationship observed in practice could be an artefact, and hence is likely to be unstable. It would also suggest that the relationship may depend more on the specific variant used rather than the technique itself.

Evidence

The extent and significance of the correlation between SG, TTO and the VAS has been reported in an earlier section. However, correlation is a poor measure of agreement. In over 30 studies reporting comparisons of the VAS and SG (see appendix 2) SG values consistently exceeded those of the VAS (e.g. Torrance, 1976; Bombadier *et al.*, 1982; Llewellyn-Thomas *et al.*, 1984; Read *et al.*, 1984; Bass *et al.*, 1994). Only 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 and Sutton, 1997). The 18 studies reporting TTO and the VAS results found a less consistent relationship, nonetheless, in the majority of studies TTO values exceeded those for the VAS.

Studies mapping the relationship between the VAS and the choice-based methods have used regression analyses. Some use aggregate level data sets, where the regression analyses have been undertaken on mean health state values, and others have used individual level data. It has been argued that since the theories operate at the individual level, then this is the appropriate level of analysis (Loomes, 1993; Dolan and Sutton, 1997). Others have taken a more pragmatic line, arguing that since the results are going to be used in cost–utility analyses of mean results, where an aggregate model is more appropriate (Stiggelbout *et al.*, 1996). We review both levels of analysis.

VAS to SG

Five published studies have been found with empirical estimates of the relationship between SG and the VAS (Torrance *et al.*, 1982; Bombardier *et al.*, 1982; Torrance *et al.*, 1996; Bleichrodt and Johannesson, 1996; Dolan and Sutton, 1997) and a conference paper (Loomes *et al.*, 1994). Torrance *et al.* (1982) first published the power relationship shown in equation (1) with its power term equalling 1.61. However, this was not directly estimated

from VAS and SG data. The equation was estimated on VAS and TTO data (Torrance, 1976) and they justified its extension to SG because the earlier study found TTO values were equal to SG (a finding contradicted by most subsequent studies).

The first published study using SG and VAS data was by Bombardier *et al.* (1982), who estimated a linear relationship between 35 mean VAS and SG values explaining 76% of the variation. The constant term was found to be highly significant. Torrance *et al.* (1982) and Loomes (1993) have subsequently fitted a power function to the same data with an estimated value for the power term of 2.16 and 2.27 (and explaining 80% of the variation), respectively. Torrance and colleagues estimated weights for the HUI-II using the relationship between mean SG and VAS values for four health states. They only fitted a power function, and this achieved a 97% fit and a value of 2.29 for the power term. None of these studies presented standard econometric diagnostic information on the specification of the models. It is therefore not possible to compare the models or judge how well they fitted the data.

At the individual level, Dolan and Sutton (1997), Bleichrodt and Johannesson (1996) and Loomes *et al.* (1994) have been able to estimate a power relationship. The former study used a Tobit procedure rather than conventional ordinary least squares to allow for the censored nature of the data. The power function was, however, outperformed by the linear models in terms of specification and ability to explain variations in the data. Furthermore, the power function failed all diagnostic tests of the model. The second study only estimated power functions, but also found evidence of heterogeneity and, in many cases, of misspecification. These studies also found more serious problems with the relationship. Bleichrodt and Johannesson (1996) and Loomes *et al.* (1994) found the parameter estimate of the power term (i.e. *b*) varied depending on the context in which the health states values were elicited by the VAS. Dolan and Sutton (1997) also found that different variants of SG (i.e. props and no props) altered the parameter values of the models.

VAS to TTO

There were four published studies modelling the empirical relationship between the VAS and TTO found by the search (Torrance, 1976; Bombardier *et al.*, 1982; Stiggelbout *et al.*, 1996; Dolan and Sutton, 1997) and an unpublished thesis (van Busschbach, 1994). Three of these studies have already been reported. Torrance's estimated a power function for TTO to the VAS from 18 pairs of health state values able to explain 79% of the variation and with a power term of 1.61. Bombardier *et al.* (1982) estimated a linear model for the relationship between the VAS and TTO for the 35 pairs of mean values, explaining 89% of the variation. Loomes (1993) estimated a power function with 1.82 for the power term and 88% of variation. Again it is not possible to compare these models. Stiggelbout *et al.* (1996) have claimed to replicate Torrance's original findings on 183 cancer patients rating their own health. They estimated a power function able to explain 72% of the data and a power term of 1.55. They did not examine other functional forms, nor did they report any diagnostics. Stiggelbout *et al.* (1994) cited the unpublished work of van Busschbach, who found the power model was no better than the linear model, and estimated the power term to be 2.13.

Neither Torrance (1976) nor Stiggelbout *et al.* (1996) were able to fit satisfactory power functions to data at the individual level. Only Dolan and Sutton (1997) seem to have done this. However, the power models were again outperformed by the linear ones, and the power models suffered from heterogeneity and misspecification. The mapping functions were also found to differ substantially between the two variants of TTO.

Discussion

The purpose of this section has been to examine whether it is possible to estimate unique and stable relationships between the VAS and SG and the VAS and TTO. The amount of evidence available to address this question was limited. There were only seven studies in total with statistically estimated relationships (with two studies looking at SG and TTO). Furthermore, these differed in the level of analysis, the functional forms examined and the diagnostic testing performed. However, some indications of the likelihood of finding sufficiently stable relationships can be gleaned.

At the individual level, only one out of three studies were able to estimate a statistical model for TTO and three for SG. The evidence on TTO and SG, such as it was, could not distinguish between the theories described above. There was no evidence of a power function performing better than a linear form, and indeed where they were formally compared the latter performed better in statistical terms. This may indicate that more than one explanation operates. More larger studies are required to address the question. It has also been suggested that a better way of assessing what

underlies respondents' answers to the different valuation techniques would be to conduct interviews (Loomes *et al.*, 1994; Robertson *et al.*, 1995).

For the power functional form, the value of the power term is reported to vary from 1.55 (Stiggelbout *et al.*, 1996) to 2.13 (van Busschbach, 1994) for TTO and 1.59 (Bleichrodt and Johannesson, 1996) to 6.41 (Loomes *et al.*, 1994) for SG. These values for *b* translate into differences in health state values of up to 0.11 for TTO and 0.28 for SG on a 0 to 1.00 scale. These are likely to have considerable implications for the incremental cost per QALY of different interventions.

Variations in model parameters may reflect differences in attitudes to risk or time, the role of different reference points or simply a more complex relationship than has been modelled in these studies. Some of these differences may have arisen from the use of different variants of the techniques, such as whether or not death and full health were used as end-points in the VAS, which has been suggested by Stigglebout *et al.* (1996), or the severity of the other states being valued in the VAS task, which was found by Loomes *et al.* (1994) and Bleichrodt and Johannesson (1996) to be associated with different parameter values. 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 seems no straightforward way of converting one into the other which is stable across procedures and contexts'.

Models for TTO and SG performed better at the group level, and were able to explain a majority of the variation. On pragmatic grounds Stigglebout *et al.* (1996) have argued that this is more appropriate for cost–utility analysis. Furthermore, less variation has been found in the parameter values of the VAS to SG relationship at this level. Existing published evidence is limited to two studies for SG and three for TTO. These were based on very small numbers in some instances (e.g. Torrance *et al.* (1996) used four pairs of mean health state values), and none reported diagnostic test results for evaluating the appropriateness of different specifications. There has also been no attempt to test the ability of these models to predict directly obtained mean SG values on an independent data set. This lack of evidence would seem to provide an insecure foundation on which to conclude that there is a unique and stable relationship between the VAS and SG or TTO. We therefore recommend SG and TTO values are obtained directly, rather than estimated from VAS values, until better evidence becomes available at the group level.

Conclusions

Considering the available literature as a whole, several observations can be made. We have considered techniques against the specified review criteria and we have considered the mapping of values from the VAS to SG and TTO, and conclude the following:

- **Practicality**. ME and PTO lack empirical support but there is little to choose between the other techniques discussed, all have proved to be practical on most populations, although VAS techniques have performed slightly better and have cost advantages.
- **Reliability**. There is little evidence relating to ME and PTO and there is very little difference between the performance of SG, TTO and VAS techniques. Present evidence does not offer a basis to differentiate between the SG, TTO and VAS techniques.
- **Theoretical validity**. Only choice based techniques should be used: SG, TTO and PTO. The choice between SG, TTO and PTO depends on the perspective employed. The debate surrounding SG versus TTO is unresolved, and depends on ones belief in the descriptive validity of the theory or its normative basis.
- **Empirical validity**. Empirical evidence relating to the theoretical perspective of the techniques has shown that there are problems with all techniques in terms of descriptive validity. The empirical evidence available to inform on the performance of techniques against preferences would suggest that (1) VAS techniques may be measuring aspects of health status rather than valuing health states, and (2) choice-based methods are best placed to reflect strength of preference for health states. Other than that it seems there is little to choose between the choice-based methods. SG and TTO are the most developed techniques, with PTO being relatively undeveloped as well as focusing on social preference. SG and TTO have been found to give similar results, although SG values tend to exceed those of TTO. At present the empirical literature informing on empirical validity would favour TTO, although this currently reflects the sparse literature available.
- **Mapping of values**. Given the current literature relating to the modelling of SG and TTO values from VAS responses, we conclude that the relationship between these techniques is not robust, and current evidence provides an insecure foundation for a unique and stable relationship between them. We recommend that

SG and TTO values are obtained directly, rather than trying to estimate them from VAS values.

Taking a wider view, it would seem that specific valuation techniques may be more appropriate in specific situations. The literature would lead us to believe that TTO is more suited than SG to the valuation of chronic health states lasting for a number of years (Dolan and Gudex, 1995); this may be demonstrated by the relatively large number of studies in the area of renal disease which have used the TTO to good effect (Laupacis *et al.*, 1996; Churchill *et al.*, 1990, 1991; Russell *et al.*, 1992; Molzahn *et al.*, 1996). Whereas, SG is perhaps conceptually preferable in situations where there is an actual risk, for example hip replacement (Laupacis *et al.*, 1993). Unlike TTO, SG permits the possibility of a catastrophe (death at young age), which may more closely reflect patients concerns, in certain situations, than does giving up some remaining life expectancy at a point in the future (Reed *et al.*, 1993). When using SG to value health states, which involve a very large or a very small risk, that is, less than 0.1 or greater than 0.9, evidence would suggest that respondents may overestimate or underestimate the true risks involved (Kahneman and Tversky, 1982; Loomes and Mckenzie, 1989), therefore researchers are advised to consider such effects in advance.

We would advocate a considered approach to the selection and use of health state valuation techniques. Although we have highlighted the fact that the present literature covering the valuation techniques is sparse, we feel it can offer an invaluable insight, both theoretically and empirically, to those wishing to elicit health state valuations. We support the use of SG or TTO, yet the choice between the techniques will depend upon the particular study question and characteristics. PTO would also seem to present a promising technique for the elicitation of social preferences, yet further development is still required.

Further research evidence is required to inform on the performance of techniques in terms of practicality, reliability, theoretical validity and especially empirical validity. We see SG, TTO and PTO as the techniques most deserving of further examination. More recently, researchers have turned to a more qualitative style of investigation (e.g. Robinson *et al.*, 1997). We see this as a fruitful avenue of investigation, and would advocate further qualitative research to assess health state valuation techniques.

Readers are reminded that we have not considered many of the factors which may impact upon the use of health state valuation techniques. We have only considered technique-specific issues present in the literature, and in doing so have not covered non-technique sources of variation, such as method of administration (e.g. postal survey or interview-based methods of administration), reference effects (e.g. which health states to use and in which order) or whose values to elicit (e.g. patients, physicians, public). For further guidance on these issues, readers are advised to consult Froberg and Kane (1989a,b) and Dolan and Sutton (1997) as a starting point.

References

A full list of references, identified through the systematic search procedure, is provided in appendix 2.

Ashby J, O'Hanlon M, Buxton MJ. The time trade-off technique: how do the valuations of breast cancer patients compare to those of other groups? *Q Life Res* 1994;**3**:257–65.

Bakker C, Rutten M, van Doorslaer E, Bennett K, van der Linden S. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. *J Rheumatol* 1994;**21**:269–74.

Bakker C, Rutten M, van Santen Hoeufft M, Bolwijn P, van Doorslaer E, Bennett K, *et al.* Patient utilities in fibromyalgia and the association with other outcome measures. *J Rheumatol* 1995;*22*:1536–43.

Bass EB, Steinberg EP, Pitt HA, Griffiths RI, Lillemoe KD, Saba GP, *et al.* Comparison of the rating scale and the standard gamble in measuring patient preferences for outcomes of gallstone disease. *Med Decis Making* 1994;**14**:307–14.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.

Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *J Health Econ* 1997a;**16**:155–75.

Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. *Med Decis Making* 1997b;**17**:208–16.

Bombadier C, Wolfson AD, Sinclair AJ, McGreer A. Comparison of three measurement technologies in the evaluation of a functional status index. In: Deber R, Thompson G, editors. Choices in health care: decision making and evaluation of effectiveness, Toronto: University of Toronto, 1982.

Bosch JL, Hunink MG. The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication. *Med Decis Making* 1996;**16**:217–25.

Boyd NF, Sutherland HJ, Heasman KZ, Tritchler DL, Cummings BJ. Whose utilities for decision analysis? *Med Decis Making* 1990;**10**:58–67.

Broome J. QALYs. *J Public Econ* 1993;**50**:149–67.

Buckingham JK, Birdsall J, Douglas JG. Comparing three versions of the time tradeoff: time for a change? *Med Decis Making* 1994;**16**:335–47.

Buckingham K. A note on HYE (healthy years equivalent). *J Health Econ* 1993;**12**:301–9.

Busschbach JJ, Horikx PE, van den Bosch JM, Brutel de la Riviere A, de Charro FT. Measuring the quality of life before and after bilateral lung transplantation in patients with cystic fibrosis. *Chest* 1994;**105**:911–17.

Cairns J, Shackley P, Hundley V. Decision making with respect to diagnostic testing: a method of valuing the benefits of antenatal screening. *Med Decis Making* 1996;**16**:161–8.

Cher DJ, Miyamoto J, Lenert LA. Incorporating risk attitude into Markov-process decision models: importance for individual decision making. *Med Decis Making* 1997;**17**: 340–50.

Churchill D, Keown P, Laupacis A, Muirhead N, Sim D, Slaughter D, *et al.* Association between recombinant human erythropoietin and quality of life and exercise capacity of patients receiving haemodialysis. *BMJ* 1990;**300**:573–8.

Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizu A, *et al.* Measurement of quality of life in end-stage renal disease: the time trade-off approach. *Clin Invest Med* 1987;**10**:14–20.

Clarke AE, Goldstein MK, Michelson D, Garber AM, Lenert LA. The effect of assessment method and respondent population on utilities elicited for Gaucher disease. *Q Life Res* 1997;**6**:169–84.

Currim IS, Sarin RK. A comparative evaluation of multi-attribute consumer preference models. *Management Sci* 1984;**30**:543–61.

Detsky AS, McLaughlin JR, Abrams HB, L'Abbe KA, Whitwell J, Bombardier C, *et al.* Quality of life of patients on long-term total parenteral nutrition at home. *J Gen Intern Med* 1986;**1**:26–33.

Dolan P. The effect of experience of illness on health state valuations. *J Clin Epidem* 1996a;**49**:551–64.

Dolan P. Modelling valuations for health states: the effect of duration. *Health Policy* 1996b;**38**:189–203.

Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996a;**15**:209–31.

Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. *Health Econ* 1996b;**5**:141–54.

Dolan P, Gudex C. Time preference, duration and health state valuations. *Health Econ* 1995;**4**:289–99.

Dolan P, Kind P. Inconsistency and health state valuations. *Soc Sci Med* 1996;**42**:609–15.

Dolan P, Sutton M. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. *Soc Sci Med* 1997;**44**:1519–30.

Drummond MF, Davies L. Economic analysis alongside clinical trials: revisiting the methodological issues. *Int J Technol Assess Health Care* 1991;**7**:561–73.

Dyer JS, Sarin RK. Relative risk aversion. *Management Sci* 1982;**28**:875–86.

Essink Bot ML, Bonsel GJ, van der Maas PJ. Valuation of health states by the general public: feasibility of a standardized measurement procedure. *Soc Sci Med* 1990;**31**:1201–6.

Feeney DH, Torrance GW. Incorporating utility based quality of life assessment in clinical trials. *Med Care* 1989;**27**:S190–204.

Ferguson BM, Keown PA. An introduction to utility measurement in health care. *Infect Control Hosp Epidemiol* 1995;**16**:240–7.

Ferraz MB, Quaresma MR, Goldsmith CH, Bennett K, Atra E. Corticosteroids in patients with rheumatoid arthritis: utility measurements for evaluating risks and benefits. *Rev Rheum Engl Ed* 1994;**61**:240–4.

Fischhoff B, Goitein B, Shapira Z. The expected utility of expected utility approaches. In: Feather NT, editor. Expectations and actions: expectancy-value models in psychology, Hillsdale, NJ: Lawrence Earlbaum, 1982.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – II: scaling methods. *J Clin Epidemiol* 1989a;**42**:459–71.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – III: population and context effects. *J Clin Epidemiol* 1989b;**42**:585–92.

Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K, *et al.* The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making* 1993;**13**:89–102.

Gabriel SE, Campion ME, O'Fallon WM. Patient preferences for nonsteroidal antiinflammatory drug related gastrointestinal complications and their prophylaxis. *J Rheumatol* 1993;**20**:358–61.

Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health Serv Res* 1994;**29**:207–24.

Gafni A. HYEs: do we need them and can they fulfil the promise? *Med Decis Making* 1996;**16**:215–16.

Gage BF, Cardinalli AB, Owens DK. The effect of stroke and stroke prophylaxis with aspirin or warfarin on quality of life. *Arch Intern Med* 1996;**156**:1829–36.

Glasziou PP, Bromwich S, Simes RJ. Quality of life six months after myocardial infarction treated with thrombolytic therapy. AUS-TASK Group. Australian arm of International tPA/SK Mortality Trial. *Med J Aust* 1994;**161**:532–6.

Gudex C, Kind P, van-Dalen H, Durand MA, Morris J, Williams A. Comparing scaling methods for health state valuations – Rosser revisited. Discussion paper 107. York: Centre for Health Economics, University of York, 1993.

Gudex C, Dolan P, Kind P, Williams A. Health state valuations from the general public using the visual analogue scale. *Q Life Res* 1996;**5**:521–31.

Hadorn DC. Setting health care priorities in Oregon: cost-effectiveness meets the rule of rescue. *JAMA* 1991;**265**:2218–25.

Handler RM, Hynes LM, Nease RF Jr. Effect of locus of control and consideration of future consequences on time tradeoff utilities for current health. *Q Life Res* 1997;**6**:54–60.

Hershey JC, Kunrather HG, Schoemaker PJH. Sources of bias in assessment procedures for utility functions. *Manag Sci* 1981;**28**:936–54.

Hornberger JC, Redelmeier DA, Petersen J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *J Clin Epidemiol* 1992;**45**:505–12.

Irvine EJ. Quality of life in inflammatory bowel disease: biases and other factors affecting scores. *Scand J Gastroenterol Suppl* 1995;**208**:136–40.

Johannesson M. QALYs, HYEs and individual preferences – a graphical illustration. *Soc Sci Med* 1994;**39**:1623–32.

Johannesson M, Jonsson B, Karlson G. Outcome measurement in economic evaluation. *Health Econ* 1996;**5**:279–96.

Johnson ES, Sullivan SD, Mozaffari E, Langley PC, Bodworth NJ. A utility assessment of oral and intravenous ganciclovir for the maintenance treatment of AIDS-related cytomegalovirus retinitis. *PharmacoEconomics* 1996;**10**:623–9.

Jones-Lee M, Loomes G, O'Reilly D, Phillips P. The value of preventing non-fatal road injuries: findings of a willingness-to-pay national sample survey. Transport Research Laboratory contractor report 330. Transport Research Laboratory, 1993.

Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;**47**:263–91.

Kaplan RM, Bush JW, Berry CC. Health status index: category rating versus magnitude estimation for measuring levels of well-being. *Med Care* 1979;**17**:501–25.

Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures. *Q Life Res* 1993;**2**:467–75.

Kavanagh T, Myers MG, Baigrie RS, Mertens DJ, Sawyer P, Shephard RJ. Quality of life and cardiorespiratory function in chronic heart failure: effects of 12 months' aerobic training. *Heart* 1996;**76**:42–9.

Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value trade-offs. New York: Wiley, 1976.

Kind P, Rosser R, Williams A. A valuation of quality of life: some psychometric evidence. In: Jones-Lee MW, editor. The value of life and safety. Amsterdam: North-Holland, 1982:159–70.

Krahn MD, Mahoney JE, Eckman MH, Trachtenberg J, Pauker SG, Detsky AS. Screening for prostate cancer. A decision analytic view. *JAMA* 1994;**272**:773–80.

Kreibich DN, Vaz M, Bourne RB, Rorabeck CH, Kim P, Hardie R, *et al.* What is the best way of assessing outcome after total knee replacement? *Clin Orthop* 1996;**331**:221–5.

Krumins PE, Fihn SD, Kent DL. Symptom severity and patients' values in the decision to perform a trans-urethral resection of the prostate. *Med Decis Making* 1988;**8**:1–8.

Laupacis A, Bourne R, Rorabeck C, Feeny D, Wong C, Tugwell P, *et al.* The effect of elective total hip replacement on health-related quality of life. *J Bone Joint Surg Am* 1993;**75**:1619–26.

Laupacis A, Keown P, Pus N, Krueger H, Ferguson B, Wong C, *et al.* A study of the quality of life and cost–utility of renal transplantation. *Kidney Int* 1996;**50**:235–42.

Lenert LA, Soetikno RM. Automated computer interviews to elicit utilities: potential applications in the treatment of deep venous thrombosis. *J Am Med Informatics Assoc* 1997;**4**:49–56.

Lenert LA, Morss S, Goldstein MK, Bergen MR, Faustman WO, Garber AM. Measurement of the validity of utility elicitations performed by computerized interview. *Med Care* 1997;**35**:915–20.

Lipscomb J. Value preferences for health: meaning, measurement, and use in program evaluation. In: Kane RL, Kane RA, editors. Values and long term care. Lexington, Mass: Lexington Books, 1982:27–83.

Llewellyn Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, *et al.* The measurement of patients' values in medicine. *Med Decis Making* 1982;**2**:449–62.

Llewellyn Thomas HA, Williams JI, Levy L, Naylor CD. Using a trade-off technique to assess patients' treatment preferences for benign prostatic hyperplasia. *Med Decis Making* 1996;**16**:262–82.

Loomes G. Disparities between health state measures: is there a rational explanation? In: Gerrard W, editor. The economics of rationality. London: Routledge, 1993.

Loomes G, Jones Lee M, Robinson A. What do visual analogue scales actually mean? Paper presented to HESG Conference, Newcastle, 1994.

Loomes G, McKenzie L. The use of QALYs in health care decision making. *Soc Sci Med* 1989;**28**:299–308.

Loomes G, Sugden R. Regret theory: an alternative theory of rational choice under uncertainty. *Econ J* 1982;**92**:805–24.

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. Oxford: Oxford University Press, 1996.

McLeod RS, Churchill DN, Lock AM, Vanderburgh S, Cohen Z. Quality of life of patients with ulcerative colitis preoperatively and postoperatively. *Gastroenterol* 1991;**101**:1307–13.

Mehrez A, Gafni A. Evaluating health-related quality of life: an indifference curve interpretation for the time trade-off technique. *Soc Sci Med* 1990;**31**:1281–3.

Mehrez A, Gafni A. The healthy-years equivalents: how to measure them using the standard gamble approach. *Med Decis Making* 1991;**11**:140–6.

Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. *Med Decis Making* 1993;**13**:287–92.

Molzahn AE, Northcott HC, Hayduk L. Quality of life of patients with end stage renal disease: a structural equation model. *Q Life Res* 1996;**5**:426–32.

Morrison GC. Consistency within and between methods of health status valuation: a within subject examination of the willingness to pay and standard gamble methods. Econometric Society European Meeting. Maastricht, August, 1994.

Morss SE, Lenert LA, Faustman WO. The side effects of antipsychotic drugs and patients' quality of life: patient education and preference assessment with computers and multimedia. *Proc Ann Symp Comput Appl Med Care* 1993;17–21.

Murray CJ, Lopez AD. Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: global Burden of Disease Study. *Lancet* 1997;**349**:1347–52.

Nease RF Jr, Kneeland T, O'Connor GT, Sumner W, Lumpkins C, Shaw L, *et al.* Variation in patient utilities for outcomes of the management of chronic stable angina. Implications for clinical practice guidelines. Ischemic Heart Disease Patient Outcomes Research Team. *JAMA* 1995;**273**:1185–90 (erratum: *JAMA* 1995;**274**(8):612).

Nease RF Jr, Tsai R, Hynes LM, Littenberg B. Automated utility assessment of global health. *Q Life Res* 1996;**5**:175–82.

Nichol G, Llewellyn Thomas HA, Thiel EC, Naylor CD. The relationship between cardiac functional capacity and patients' symptom-specific utilities for angina: some findings and methodologic lessons. *Med Decis Making* 1996;**16**:78–85.

Nicholl J, Brazier JE, Milner PC, Westlake L, Kohler B, Williams BT, *et al.* Randomised controlled trial of cost-effectiveness of lithotripsy and open cholecystectomy as treatments for gallbladder stones. *Lancet* 1992;**340**:801–7.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Plann Manage* 1991a;**6**:234–42.

Nord E. EuroQoL: health-related quality of life measurement. Valuations of health states by the general public in Norway. *Health Policy* 1991b;**18**:25–36.

Nord E. Methods for quality adjustment of life years. *Soc Sci Med* 1992a;**34**:559–69.

Nord E. An alternative to QALYs: the saved young life equivalent (SAVE). *BMJ* 1992b;**305**:875–7.

Nord E. The trade-off between severity of illness and treatment effect in cost-value analysis of health care. *Health Policy* 1993;**24**:227–38.

Nord E. The person-trade-off approach to valuing health care programs. *Med Decis Making* 1995;**15**:201–8.

Nord E, Richardson J, Macarounas Kirchmann K. Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. *Int J Technol Assess Health Care* 1993;**9**:463–78.

O'Brien B, Viramontes JL. Willingness to pay: a valid and reliable measure of health state preference? *Med Decis Making* 1994;**14**:289–97.

Parducci A. Contextual effects. A range-frequency analysis. In: Carterette E, Friedman M, editors. Handbook of perception, vol II. New York: Academic Press, 1974.

Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Services Res* 1973;**8**:228–45.

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Med Decis Making* 1994;**14**:9–18.

Perez DJ, McGee R, Campbell AV, Christensen EA, Williams S. A comparison of time trade-off and quality of life measures in patients with advanced cancer. *Q Life Res* 1997;**6**:133–8.

Pinto Prades JL. Is the person trade-off a valid method for allocating health care resources? *Health Econ* 1997;**6**:71–81.

Provenzale D, Shearin M, Phillips Bute BG, Drossman DA, Li Z, Tillinger W, *et al.* Health-related quality of life after ileoanal pull-through evaluation and assessment of new health status measures. *Gastroenterology* 1997;**113**:7–14.

Rabin R, Rosser RM, Butler C. Impact of diagnosis on utilities assigned to states of illness. *J R Soc Med* 1993;**86**:444–8.

Ramsey SD, Patrick DL, Lewis S, Albert RK, Raghu G. Improvement in quality of life after lung transplantation: a preliminary study. The University of Washington Medical Center Lung Transplant Study Group. *J Heart Lung Transplant* 1995;**14**:870–7.

Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes. Comparison of assessment methods. *Med Decis Making* 1984;**4**:315–29.

Redelmeier DA, Heller DN. Time preference in medical decision making and cost-effectiveness analysis. *Med Decis Making* 1993;**13**:212–17.

Reed WW, Herbers JE Jr, Noel GL. Cholesterol-lowering therapy: what patients expect in return. *J Gen Intern Med* 1993;**8**:591–6.

Revicki DA. Relationship between health utility and psychometric health status measures. *Med Care* 1992;**30**:MS274–82.

Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Q Life Res* 1993;**2**:477–87.

Revicki DA, Wu AW. Discrimination and responsiveness of health status and utility measures. *Ann Meet Int Soc Technol Assess Health Care* 1995;**11**:abstract 6.

Revicki DA, Wu AW, Murray MI. Change in clinical status, health status, and health utility outcomes in HIV-infected patients. *Med Care* 1995;**33**:AS173–82.

Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994;**39**:7–21.

Robinson A, Dolan P, Williams A. Valuing health states using VAS and TTO: what lies behind the numbers? *Soc Sci Med* 1997;**45**:1289–97.

Rosser R, Kind P. A scale of valuations of states of illness: is there a social consensus? *Int J Epidemiol* 1978;**7**:347–58.

Russell JD, Beecroft ML, Ludwin D, Churchill DN. The quality of life in renal transplantation – a prospective study. *Transplantation* 1992;**54**:656–60.

Rutten van Molken MP, Bakker CH, van Doorslaer EK, van der Linden S. Methodological issues of patient utility measurement. Experience from two clinical trials. *Med Care* 1995a;**33**:922–37.

Rutten van Molken MP, Custers F, van Doorslaer EK, Jansen CC, Heurman L, Maesen FP, *et al.* Comparison of performance of four instruments in evaluating the effects of salmeterol on asthma quality of life. *Eur Respir J* 1995b;**8**:888–98.

Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chronic Dis* 1981;**31**:697–704.

Schoemaker PJH. The expected utility model: its variants, purposes, evidence and limitations. *J Econ Lit* 1982;529–63.

Scott A. Giving things up to have more of others. The implications of limited substitutability for eliciting preferences in health and health care. Paper presented to HESG Conference, Sheffield, 1998.

Sintonen H. An approach to measuring and valuing health states. *Soc Sci Med* 1981;**15C**:55–65.

Sivertssen E, Field NB, Abdelnoor M. Quality of life after open heart surgery. *Vasc Surg* 1994;**28**:581–8.

Stevens SS. A metric for social consensus. *Science* 1966;**151**:530–41.

Stevens SS. Issues in psychophysical measurement. *Psychol Rev* 1971;**78**:426–50.

Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Med Decis Making* 1994;**14**:82–90.

Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Habbema JD, de Haes JC. The "utility" of the time trade-off method in cancer patients: feasibility and proportional trade-off. *J Clin Epidemiol* 1995;**48**:1207–14.

Stiggelbout AM, de Haes JC, Kiebert GM, Kievit J, Leer JW. Tradeoffs between quality and quantity of life: development of the QQ Questionnaire for Cancer Patient Attitudes. *Med Decis Making* 1996a;**16**:184–92.

Stiggelbout AM, Eijkemans MJ, Kiebert GM, Kievit J, Leer JW, De Haes HJ. The 'utility' of the visual analog scale in medical decision making and technology assessment. Is it an alternative to the time trade-off? *Int J Technol Assess Health Care* 1996b;**12**:291–8.

Sutherland HJ, Llewellyn Thomas H, Boyd D, Till JE. Attitudes towards quality of survival: the concept of maximum endurable time. *Med Decis Making* 1982;**2**:299–309.

Sutherland HJ, Dunn V, Boyd NF. Measurement of values for states of health with linear analog scales. *Med Decis Making* 1983;**3**:477–87.

Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Econ Plan Sci* 1976;**10**:129–36.

Torrance GW. Preferences for health states: a review of measurement methods. *Mead Johnson Symp Perinat Dev Med* 1982;37–45.

Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986;**5**:1–30.

Torrance GW. Utility approach to measuring health-related quality of life. *J Chronic Dis* 1987;**40**:593–603.

Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care* 1989;**5**:559–75.

Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care* 1996;**34**:702–22.

Tsevat J, Goldman L, Soukup JR, Lamas GA, Connors KF, Chapin CC, *et al.* Stability of time-tradeoff utilities in survivors of myocardial infarction. *Med Decis Making* 1993;**13**:161–5.

Tversky A, Slovic P, Kahneman D. The causes of preference reversal. *Am Econ Rev* 1990;**80**:204–17.

Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Individual utilities are inconsistent with rationing choices: a partial explanation of why Oregon's cost-effectiveness list failed. *Med Decis Making* 1996;**16**:108–16.

Van Busschbach J. De validiteit van QALY's [The validity of QALY's]. Unpublished PhD thesis. Rotterdam: Erasmus University Rotterdam, Sanders Instituut, 1994.

van der Donk J, Levendag PC, Kuijpers AJ, Roest FH, Habbema JD, Meeuwis CA, Schmitz PI. Patient participation in clinical decision-making for treatment of T3 laryngeal cancer: a comparison of state and process utilities. *J Clin Oncol* 1995;**13**:2369–78.

Von Meumann J, Morganstern O. Theory of games and economic behavior. Princeton: Princeton University Press, 1944.

Wakker P, Stiggelbout A. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Med Decis Making* 1995;**15**:180–6.

Wolfson AD, Sinclair AJ, Bombardier C, McGreer A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA, editors. Values and long term care. Lexington, Mass. Lexington Books, 1982:191–214.

Zug KA, Littenberg B, Baughman RD, Kneeland T, Nease RF, Sumner W, *et al.* Assessing the preferences of patients with psoriasis. A quantitative, utility approach. *Arch Dermatol* 1995;**131**:561–8.

Chapter 5

A review of five MAUSs

MAUSs are widely used in economic evaluations alongside clinical studies to value the benefits of health care in terms of QALYs. The crucial difference between these and other measures of HRQoL is that the weights used to score them have been obtained using one of the preference elicitation techniques described in chapter 4 (Drummond *et al.*, 1987). Furthermore, MAUSs produce a score of 1 or less, where 1 is equivalent to full health and 0 is death. Scores take a negative value for health states regarded as worse than death.

Five MAUSs have been found from a search of the literature to have been used in studies published up to the end of 1996, and these are the instruments reviewed in this chapter. The five are: the QWB, Rosser's disability/distress classification, the HUI (mark I to III), the EQ-5D and the 15D. These MAUSs differ considerably in terms of their dimensions, size, and methods of valuation, and they have been developed in different countries (*Table 2*). There is little guidance in the literature on which MAUSs to use, and there has been no systematic review of these scales against economic criteria.

The MAUSs are reviewed in this chapter in terms of the criteria of practicality, reliability and validity using the check-list for judging preference-based measures developed in chapter 3. The review is conducted on papers identified through a systematic search of the literature. The chapter begins by briefly describing the characteristics of each instrument. This is followed by a description of the methods and results of the systematic search. The

TABLE 2 *Characteristics of MAUSs*

MAUS	Descriptive characteristics			Valuation characteristics			
	Dimension	**Levels**	**Health states**	**Valuation technique**	**Method of extrapolation**	**Sample**	**Country**
Rosser classification	Disability Distress	8 4	29	(1) ME (2) Synthesis of ME, VAS, TTO	None None	70 (selected) 140 (representative)	UK (London)
QWB	Mobility, physical activity, social functioning 27 symptoms/problems	3 2	1170	VAS	Modelling	866 (representative)	USA (San Diego)
HUI-II	Sensory, mobility, emotion Cognitive, self-care, pain Fertility	4–5 3	24,000	VAS transformed into SG	Algebraic	203 (parents)	Canada (Hamilton)
HUI-III	Vision, hearing, speech Ambulation, dexterity Emotion, cognition, pain	5–6	972,000	VAS transformed into SG	Algebraic	504 (representative)	Canada (Hamilton)
EQ-5D	Mobility, self-care, usual activities, pain/discomfort Anxiety/depression	3	243	MVH – TTO and VAS	Modelling	3395 (representative)	UK
15D	Mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort/symptoms, depression, distress, vitality, sexual activity	4–5	Billions	VAS	Algebraic		Finland

bulk of the chapter is devoted to a detailed review of each instrument. The final section compares the instruments on the basis of the results of the review.

Description of the instruments

Quality of Well-Being Scale

The 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, though there have been a number of revisions to its wording, its size and the preference weights (Patrick *et al.*, 1973a; Kaplan *et al.*, 1976; Bush *et al.*, 1982; Kaplan and Anderson, 1988; Kaplan, 1993a). 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.

The HSC contains two components (*Table 2*): three multilevel dimensions relating to function (mobility, physical activity and social activity) and a list of 27 symptom and problem complexes (e.g. 'general tiredness, weakness, or weight loss', 'wore eyeglasses or contact lenses'). The functional dimensions and the symptom/complexes combine to form 1170 health states. A patient is assigned to this classification from an interview.

An overall health state score is calculated by a simple additive formula, that is, one plus the decrement (i.e. negative weight) associated with the level of each of three functioning dimensions and the most highly weighted symptom/ problem suffered by the patient.* These weights were estimated statistically from a sample of health states valued using a version of the VAS on a representative sample of the general population of San Diego, USA (Kaplan, 1989).

Rosser disability/distress scale

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

The classification has two dimensions, disability and distress, with eight and four levels, respectively. The disability dimension has descriptions for each level, for example level 3 is 'Severe social disability and/or slight impairment of performance at work. Able to do all housework except very heavy tasks', whereas the four distress levels are simply none, mild, moderate and severe. Together the two dimensions define a total of 29 health states (the matrix defines 32 states, but the worst level of disability is unconsciousness, and hence there is no distinction between the four states defined by the different levels of distress). Patients were originally classified by clinician assessment. More recently a self-completed instrument called the Health Measurement Questionnaire (HMQ) has been developed for classifying patients (Gudex and Kind, 1988). Patients have also been mapped on to the Rosser classification from other health status questionnaires (Coast, 1992).

The most commonly used weights were obtained by Rosser and her colleagues from 70 respondents using a version of ME (Kind *et al.*, 1982). The classification has since been revalued by a larger, general population sample using ME, the VAS and TTO (Gudex *et al.*, 1993). These methods produced different values, and the authors recommend a matrix of weights based on a synthesis of the results.

Health Utility Index

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

The 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. The HUI-III is an adaptation of the 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,

* $W = 1 + (CPXwt) = (MOBwt) + (PACwt) + (SACwt)$, where W is the health state score, CPX is the worst symptom/problem, MOB is the mobility scale, PAC is the physical activity scale and SAC is the social activity scale (Kapaln, 1989).

whilst 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 15-item self-completed questionnaire, from face-to-face interview and by telephone.

A utility value is obtained by inputting weights for each dimension into a multiplicative formula. These weights were estimated from valuation data obtained from a sample of parents from Hamilton, Ontario, using VAS responses transformed into SG values using a specially estimated power function. The weights for the HUI-III had not been published at the time of writing.

15D

This measure originally had a 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 the 15D is reported here, since the 15D.1 is sufficiently similar to its successor to be relevant.*

The dimensions of the 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 five levels and hence the classification is able to define many billions of health states. 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.

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. These weights were elicited from a sample of the Finnish population using versions of the VAS and ME.

EQ-5D

This instrument was developed by a multi-disciplinary 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. Patients are classified on to the EQ-5D by completing a five-item questionnaire, suitable for self-completion or interviewer administration.

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 (van Agt *et al.*, 1994; Badia *et al.*, 1995; Selai and Rosser, 1995). However, the most significant valuation work with the EQ-5D has been a large-scale survey undertaken in the UK by the MVH Group at York. Their work produced weights for valuing the EQ-5D based on TTO and VAS valuations. The result 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.

Instruments excluded from this review

This review has not included all multi-attribute utility instruments. These are briefly described below, and the reasons for their exclusion explained.

Index of Health Related Quality of Life

The Index of Health Related Quality of Life (IHQL) was developed by Rosser and Colleagues from the disability/distress classification (Rosser *et al.*, 1992; Rosser *et al.*, 1993). In the first stage of its development, distress was subdivided into physical discomfort and emotional distress. This 'three-dimensional' version defines 175 composite health states. The three dimensions have been further divided into seven attributes, and these in turn into 44 scales. The scales have been divided into 107 descriptors, which in total have 225 levels. This hierarchical classification of the IHQL defines many millions of states. The three dimensions has been valued using SG and a matrix of health state values published in an edited volume (Rosser *et al.*, 1992). The IHQL was valued using the VAS, and provisional results presented in the same volume.

Descriptions of the methods of valuation have not been published elsewhere, and it has not been

* 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).

possible to critically review this work on the basis of what is available. No applications been found in refereed journals from an extensive search of the literature.

The Australian multi-attribute utility instrument

Hawthorne, Richardson and others at the Universities of Melbourne and Monash have developed a multi-attribute utility designed for use in prioritising healthcare spending in Australia. On the basis of a literature review, consultations with health professionals and extensive psychometric testing the team have derived a classification with five 'major' dimensions (illness, independent living, social relationships, physical senses and psychological well-being) with 15 items. This developmental work has been reported in a discussion paper. The team have undertaken a valuation of health states by TTO and estimated a scoring algorithm for estimating a single index using an algebraic approach. The results of this work were not available when this report was being prepared. This is potentially an important instrument but it is not possible to undertake an extensive review at the moment.

SF-6D

A team of researchers at the University of Sheffield has developed a method for deriving a single index value from the SF-36 (Brazier *et al.*, 1998). An HSC was derived from the SF-36 that would be amenable to valuation. The result was a six-dimensional classification with between two and six levels, the SF-6D HSC. A total of 9000 health states are defined by this classification. A survey was undertaken to value a sample of states by 165 respondents including patients, health professionals, managers and students. They were each asked to value 15 health states using the VAS and SG techniques. A scoring algorithm has been estimated by statistical techniques. This work has been followed by a much larger study based on a survey of 600 UK residents representative of the UK population due to report in April 1999.

The potential of the SF-6D derives from its use of the SF-36. This has two advantages. Firstly this instrument has been shown to be more sensitive than the Rosser and EQ-5D instruments for some common conditions (see review below). The second advantage is that the SF-36 has become one of the most widely used general measures of health-related quality, and the SF-6D provides a way of deriving a preference-based single index for use in economic evaluation. However, the results were not published during the time frame of this review.

Search strategy

This review is based on a systematic search of the literature up to the end of 1996. Two approaches were used to identify articles on the MAUSs: (1) using all permutations of the names of specific scales or instruments and (2) performing an author citation search on the original articles that describe the development of each scale or instrument. The search terms are set out in *Box 4*.

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

A total of 163 papers have been identified by this strategy (and are listed by instrument in appendix 3). These were retrieved and form the material for this review. The papers were divided

BOX 4 Search strategy

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

into methodology and applications. The methodological papers (n = 92) described the instrument, the development of its classification, the derivation of its values, or provided an overview of how it can be used. These included the empirical work conducted to derive the descriptive classification and its weights. The papers reporting applications of the instrument (n = 71) provided the empirical evidence for this review. To assist in summarising these papers, the studies have been tabulated by instrument (see appendix 3). These tables describe whether the study includes evidence on the following: the patient group, the number of patients, time to complete the questionnaire, response rates, completion rates, reliability (inter-rater and retest), content and face validity, construct validity, and empirical evidence on relationship to hypothetical or stated preferences (there were no RP data).

Review of five MAUSs

Quality of Well-Being Scale

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 (see appendix 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 1 and 2 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% (Andresen *et al.*, 1995), but 100% was achieved in the study of patients with chronic obstructive pulmonary disease (COPD) (Kaplan *et al.*, 1989). The rate of completion was 93 and 100% in each of these studies respectively. Andresen *et al.* (1995) found it was more complex than the SIP and the SF-36, and Wu *et al.* (1990) and Bombardier and Ramboud (1991) also found it a complex instrument to use.

Reliability

The only published article reporting on retest reliability was an assessment of the interday reliability (Anderson *et al.*, 1989). The authors used the results of five empirical studies which found that assessments 1 day apart had correlations of 0.78–0.99 and the majority were in excess on 0.9. However, the ability of this study to assess retest 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 per cent 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).

Descriptive validity

Content and face validity. 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 US 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,* 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,

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

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 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 subscales. The list of symptoms and problems 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 important 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 achieves 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.

There has also been rewording 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 the QWB seems straightforward and in most cases reasonably clear. Some items 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 25 studies listed in appendix 2 were found to report results on the construct validity of the QWB. The QWB has been found to be significantly correlated with the general HSMs 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 effect scale), though Kaplan *et al.* (1995) found it was significantly correlated with the Beck Depression Inventory.

Holbrook *et al.* (1994) found the overall QWB score significantly improved in trauma cases between discharge and a 3 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, that is, 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 with a control group. Differences were found in the Beck Depression Inventory at 1 year, but these were not reflected in the QWB.

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* using a version of the VAS. Respondents were asked to place each state into one of 15 numbered slots defined by a scale from 0 to 16, where 0 was death and 16 optimum health.† The results were transformed on to a 0 (death) to 1 (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 made 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 the 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 the 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 the VAS can reflect preferences (see chapter 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 more recent model (Kaplan *et al.*, 1976; 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.

Empirical validity

Out of the studies listed in appendix 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 *et al.* (1993), 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

* 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).

† As described by Patrick *et al.* (1973) in an earlier publication.

of –0.36 between the number of conditions and the 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 0. 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.

Overview – QWB 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 retest 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 the 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.

Rosser classification of illness states

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 to patients, though two were reporting results from the same study (see appendix 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 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 and 95%. Completion rates were 87 and 95.5% respectively in the two studies reporting them (Hollingworth *et al.*, 1996; Kind and Gudex, 1994), but in a number of other studies the completion was 100% by implication.

Reliability

In the initial work with the classification, interclinician agreement was high (Rosser and Watt, 1972). This result was repeated with ward nurses (Benson, 1978). In a more recent study by Bryan *et al.* (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 retest reliability of results generated by the HMQ.

Questions have been raised about the assignment of patients on to 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.

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 4 is unambiguously better than 5. Gudex *et al.* (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 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 classification 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 on to the Rosser classification has also been questioned by Bryan *et al.* (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/her usual activities. The mapping of patients on to 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 classification 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 the SF-36 and EQ-5D (Hollingworth *et al.*, 1995). Furthermore, it was unable to show the improvements at 6 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 classification was found to correlate with the 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 the distress dimension, 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.

Valuation

Published work using the Rosser classification 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 ME. 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 retest and between observers (Rosser and Kind, 1978). The results were averaged across all 70 respondents and transformed on to a scale from 0 to 1, where 0 was set at death and 1 at 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). ME 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 for 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 classification choose between the original ME matrix, a new ME matrix, or a matrix based on a 'synthesis' of the 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.

Empirical validity

The studies showing the ability of the Rosser classification 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 with 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 6 months found by the EQ-5D.

Nord *et al.* (1993) compared the values of the original Rosser matrix to the responses to PTO questions. Along with the QWB and the HUI-I, it was mapped on to two EQ-6D health states. The Rosser classification 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.

Overview – Rosser 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 which 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.
- There is evidence on hypothetical preferences in group comparisons, but insensitivity was found, caused by the scoring algorithm.

Health Utility Index

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 (MAUT) to the classifications to derive the algorithms for valuing all health states. Two papers were concerned with the HUI-I, four with the HUI-II and five with the HUI-II and HUI-III. There were ten empirical studies using one of the HUI classifications (see appendix 3). HUI-II has been the most widely used to date, with seven papers. Eight of the ten 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 of McMaster University.

Practicality

The 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 and 100% and completion between 96 and 100% (see appendix 3). 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).

Reliability

In terms of inter-rater reliability, discrepancies were found in the assignment of patients on to the HUI-II, though these usually involved one dimension level (e.g. 39% disagreement was found by Feeny *et al.* (1993) and 30% by 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 test–retest 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 being 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 retest reliability (12–49 days apart) of a provisional overall HUI-III index score was found to be 0.77 (intraclass correlation coefficient).

Descriptive validity

Content and face validity. The 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 authors argue the HUI-II 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 is 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 measure, 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 the 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 the HUI-II are focused on single attribute, and in most cases reasonably short. The exception to this is emotion, where the items include a listing of moods, for example '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 the 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 the 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 the HUI-II to survivors of childhood cancer. Among 50 patients who had acute lymphoblastic leukaemia 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 with 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 ten brain tumour patients, differences were found compared with 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 acute lymphoblastic leukaemia 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 the 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).

Valuation

The HUI-II was valued by random samples of 203 parents of schoolchildren (Torrance *et al.*, 1992). Torrance and his coworkers 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 rate in the survey was 72%, 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 29% of respondents. The HUI-III has been valued by a representative sample of 504 adults from Hamilton, Ontario.

The HUI-II was valued by a random sample of parents of schoolchildren 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). The HUI-III has been valued using a stratified random sample of 504 individuals in Hamilton.

The HUI-II has been valued using a transformation of VAS ratings to SG using a power function originally estimated between the VAS and TTO. 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 4). 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 the quadratic and cubic linear models were found 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 arising 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 four states; see Torrance *et al.*, 1992)). Finally, there are major theoretical doubts about whether attitude to risk is the only difference between the VAS and SG. As reported in chapter 4, there are also doubts as to whether the VAS can be regarded as anything more than an indicator of ordinal preferences.

An important feature of the HUI has been the application of 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* has been used to value the HUIs. The multiplicative function permits a very limited form of interaction between dimensions which assume the interdependency to be the same between all dimensions and for all levels of each dimension. For the HUI-III the plan is to estimate the less restrictive multilinear functional form.

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 the 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.

Empirical validity

The HUI-II and HUI-III have not been widely used, and the only evidence on empirical validity concerned the use of the HUI-I.

Overview – HUI key points

- The 15-item questionnaire is brief and easy to use. There is no evidence on retest reliability in patient groups. The same method of administration must be used to undertake comparisons.
- The HUI-II and HUI-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 lesser extent the HUI-III, reflect concerns with the health of children.
- Applications have been very limited to date (mainly the HUI-II on survivors of childhood cancer). There is some suggestion of possible insensitivity in the HUI-II.
- The validity of the methods of valuation depends on a transformation of the VAS to SG and the unproven predictive properties of MAUT.
- There was no evidence (for or against) its empirical validity.

15D

Published literature

The search identified just nine publications, including six refereed articles, a book chapter and two working papers (see appendix 3). Five of these publications were concerned with methodology, one with the 12D (Sintonen, 1981), two with the 15D.1 (Sintonen and Pekurinen, 1993; Sintonen, 1989) and two with the 15D.2 (Sintonen, 1994a,b). All four applications have used version I of the 15D (see appendix 3). These have been supplemented by four unpublished studies described in reviews of the instrument (Sintonen and Pekurinen, 1993; Sintonen, 1994a).

Practicality

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

---

* **Types of MAUT models (Torrance *et al.*, 1995)**

Additive:

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

where

$$u(X) = \sum_{j=1}^{n} k_j = 1$$

Multiplicative (see note):

$$u(X) = \frac{1}{k} \prod_{j=1}^{n} \{[1 + k k_j u_j(x)] - 1\}$$

where

$$1 + k = \prod_{j=1}^{n} (1 + k k_j)$$

Multilinear:

$$u(x) = k_1 u_1(x_1) + k_2 u_2(x_2) + \ldots + k_{12} u_1(x_1) u_2(x_2) + k_{13} u_1(x_1) u_3(x_3) + \ldots + k_{123} u_1(x_1) u_2(x_2) u_3(x_3) + \ldots$$

where the sum of all ks equals 1. $u_j(x_j)$ is the signal attribute utility function for attribute j, $u(x)$ is the utility for health state x, represented by an n-element vector, k and k_j 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.

rates were 100% in hospital and 87% by post. Completion rates have been between 96 and 99%.

Reliability

In an unpublished study of patients waiting for coronary artery bypass grafts, the differences by dimensions between test and retest at 3 months were found to be –0.05 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 6 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.

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 the 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 unreported factor analysis, changes were made to the dimensions and their levels to form the 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 years old) had a lower score on every dimension of the 15D.2 ($p = 0.001$) 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, postoperative 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, 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 Hamilton Depression Rating Scale score was more than 16 or not 77% of the time compared with 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 the 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 the 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 versus. 25.6%), pain (21.8 versus. 14.4%) and social participation (21.8 versus 12.6), but the same for mental health and slightly better in working (8.7 versus 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.

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 VAS, as used by the EuroQoL group, and ME. 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 (½ 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.

In the 15D a decompositional approach was chosen because it would not have been possible to directly value 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.

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 postsurgery groups (Rissanen *et al.*, 1995). The prospective study of patients receiving hip and knee replacements found significant improvements 6 months after surgery. The average 15D score in coronary bypass candidates was also found significantly to improve between baseline and 3 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.

Overview – 15D key points

- 15D is a brief and easy-to-use self-completed questionnaire.
- There is some evidence of retest 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.

EQ-5D

Published literature

The search identified 40 publications, including refereed articles in journals, chapters of books, research reports and conference papers (see appendix 3). The 'grey' literature has been particularly important for the EQ-5D as this instrument is comparatively recent, 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 Medical Research Council report and a conference paper, and one published application of the EQ-6D (see appendix 3). Two of the papers

were found to be irrelevant for this review and so are not considered further.*

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.

Reliability

Three studies have examined the retest reliability of the EQ-5D: one in a sample of elderly women aged 75 years 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,b; Hurst, 1996). In the first two, the correlations between the test and retest 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 ±0.05. The reliability coefficient in the rheumatoid arthritis patients was 0.55. In all studies, these results compared well with the other generic and condition-specific health measures.

Descriptive validity

Content and face validity. The original instrument was developed from a review of other HSMs, including the QWB, SIP, NHP and the Rosser classification (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 five 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 the EQ-5D and other measures of health (the Rosser classification, NHP, QWB and SIP), that is, 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 with 26.9% for the Rosser classification, 49.1% for the SIP, 58.6% for the NHP and 58.6% for the QWB.

* The trial of treatments of menorrhagia by Sculpher and colleagues (1993) did not use the descriptive part of the 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.

The face validity of the EQ-5D has been criticised for having only three categories per dimension, which is thought to be too insensitive for detecting smaller changes (McDowell and Newell, 1996). A high proportion of respondents been classified in the top category, that is, recording no problem (Brazier *et al.*, 1993; Hollingworth *et al.*, 1995). In a general population survey using the 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 with 37–72% for the SF-36 (Brazier *et al.*, 1993). The 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 *et al.* (1993), patients who responded as having no health problem on dimensions of the EQ-6D were subdivided 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 general practitioner recently, attended an outpatient department in the last 3 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 *et al.* (1995) studied a group of patients referred for magnetic resonance imaging (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 migraine sufferers 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 versus 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 with several dimensions of the 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 6 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).

Valuation

The MVH survey was based on a large sample (n = 3395), broadly representative of the UK population (in terms of a range of sociodemographic, 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 and Thomson, 1992; Dolan *et al.*, 1996). The quality of data in terms of completeness and consistency was impressive and has been well documented (MVH Group, 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 (Dolan *et al.*, 1995). 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).

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, MVH Group, 1994).

The single index derived from the EQ-5D using the interim tariff has been found to distinguish between the general population and COPD patients (Harper *et al.*, 1995), migraine sufferers (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 versus 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 health. The EQ-5D score improved in patients who had been for a knee scan over a 6 month period, before and after reconstruction in patients with vascular disease 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 the 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.

Overview – EQ-5D key points

- It is a very brief and easy-to-use instrument.
- There is evidence of its retest 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 that the EQ-5D is able to detect large differences in line with expected preferences, though there is some contrary evidence against patient-perceived health.

Comparison of measures

The aim of this literature review was to undertake a comparison of the five MAUSs against the criteria of practicality, reliability and validity using the criteria developed in chapter 3. The applications found in the search represent a large body of work, but in terms of the range of conditions and treatments it was quite narrow.* Furthermore, there have been very few applications of the measures on the same patient populations. This limits the ability to compare the measures, since the evidence is confounded by differences in the medical condition of the patients, the treatments they receive and their sociodemographic backgrounds. Furthermore, there is far more evidence on some measures than others: the most commonly used was the QWB, followed by the Rosser classification, the EQ-5D, the HUI and the 15D. It is important to bear these problems in mind in the comparison which follows.

Practicality

All measures use a short list of questions which can be self-completed 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. There was little to choose between the questionnaires on the basis of practicality except in so far as the QWB does not have an accepted method of self-completion.

Reliability

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

Descriptive validity

The descriptive content of the measures differ widely. The size varies between the Rosser

* It is also interesting to note that the majority of studies were published by the developers of the instrument. There has been remarkably little work by independent researchers to examine the properties of these measures.

classification, with just two dimensions, compared with 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. the HUI) or actual behaviour and performance (e.g. the 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 none can be judged as better than the others in all circumstances. The exception is the Rosser disability and distress scale, which is 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 the 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.

MAUSs 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 was limited to correlations with related HSMs, with very little detailed scrutiny of the descriptive classification, whereas evidence for the HUI-II 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.

Valuation

The QWB, Rosser classification and the 15D can be regarded as inferior to the other two measures owing to their use of the VAS and ME to value the health descriptions. The HUI-II and HUI-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 the VAS on the basis of a power function which has been criticised on both theoretical and empirical grounds. The valuation of the HUI has been obtained from a smaller and less representative sample of the general population than the MVH survey. The virtues of the algebraic approach used by the HUI versus statistical methods used to value the EQ-5D has not been addressed in the literature.

Empirical validity

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.

Nord *et al.* (1993) mapped the QWB, HUI-I and Rosser classification on to EQ-6D health states. The Rosser classification was found to generate values nearest to the PTO valuations of the states, with the QWB suffering from an alleged compression towards the middle, and the HUI-I- and VAS-valued EQ-6D had much lower values. This is an interesting comparative study, but the authors recognise a number of methodological weaknesses in terms of the reliance on mapping procedures, and small samples, and the illogical ordering of the two EQ-6D instruments by the PTO technique.

Conclusions

The review has not identified one measure that is dominant across all criteria. This is in part due to the lack of empirical evidence. Future research should seek to address this problem (this is discussed in chapter 8). It is also due to differences in the content of the instruments, and the choice depends on what health changes are being measured. However, it is possible to recommend that the Rosser classification is not used in future research, given its limited coverage, the evidence of its insensitivity and the concerns about the basis of its valuation matrix. Furthermore, its main advocates in York now mainly use the EQ-5D. The QWB has the advantage of having been widely used, at least in the USA, but has been valued by the VAS, and as such is unlikely

to generate values reflecting preferences on a cardinal scale. The 15D has the largest classification, but has been the least used of the measures and also uses forms of the VAS and ME for deriving weights.

This review concludes that the best preference-based measures at the moment would seem to be the EQ-5D and the HUIs. The HUI-II and HUI-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. HUI valuations are based on a VAS ratings transformed into SG 'utilities' compared with the EQ-5D directly elicited TTO valuations. The EQ-5D has been valued by a far larger sample of the general population. Finally, they have used different means of estimating weights from the valuation of a sample of states.

We conclude that the best of the five MAUSs reviewed are the EQ-5D and the HUI. For the HUI, there is a further choice between versions depending on whether the population is children (i.e. the HUI-II) or adults (i.e. the HUI-III). This conclusion would have to be reappraised when (Canadian) weights become available for the HUI-III. We recommend research be undertaken into the validity of the descriptions in their HSCs, including comparative studies on different patient populations, and the validity of the methods for valuing them.

References

Anderson GM. A comment on the index of well-being. *Med Care* 1982;**20**:513–15.

Anderson JP, Bush JW, Berry CC. Classifying function for health outcome and quality-of-life evaluation. Self- versus interviewer modes. *Med Care* 1986;**24**:454–69.

Anderson JP, Bush JW, Berry CC. Internal consistency analysis: a method for studying the accuracy of function assessment for health outcome and quality of life evaluation. *J Clin Epidemiol* 1988;**41**:127–37.

Anderson JP, Kaplan RM, Berry CC, Bush JW, Rumbaut RG. Interday reliability of function assessment for a health-status measure – the Quality of Well-Being Scale. *Med Care* 1989;**27**:1076–84.

Anderson JP, Kaplan RM, Schneiderman LJ. Effects of offering advance directives on quality adjusted life expectancy and psychological well-being among ill adults. *J Clin Epidemiol* 1994;**47**:761–72.

Anderson RT, Aaronson NK, Wilkin D. Critical review of the international assessments of health-related quality of life. *Q Life Res* 1993;**2**:369–95.

Andresen EM, Patrick DL, Carter WB, Malmgren JA. Comparing the performance of health status measures for healthy older adults. *J Am Geriatr Soc* 1995;**43**:1030–4.

Apajasalo M, Sintonen H, Holmberg C, Sinkkonen J, Aalberg V, Pihko H, *et al.* Quality-of-life in early adolescence – a 16-dimensional health-related measure (16D). *Q Life Res* 1996;**5**:205–11.

Bakker CH, Rutten van Molken M, van Doorslaer E, Bennett K, van der Linden S. Health related utility measurement in rheumatology: an introduction. *Patient Educ Couns* 1993;**20**:145–52.

Balaban DJ, Sagi PC, Goldfarb NI, Nettler S. Weights for scoring the quality of well-being instrument among rheumatoid arthritics. A comparison to general population weights. *Med Care* 1986;**24**:973–80.

Barr RD, Furlong W, Dawson S, Whitton AC, Strautmanis I, Pai M, *et al.* An assessment of global health status in survivors of acute lymphoblastic leukaemia in childhood. *Am J Pediatr Hematol Oncol* 1993;**15**:284–90.

Barr RD, Pai MKR, Weitzman S, Feeny D, Furlong W, Rosenbaum P, *et al.* A multi-attribute approach to health status measurement and clinical management – illustrated by an application to brain tumors in childhood. *Int J Oncol* 1994;**4**:639–48.

Barr RD, Feeny D, Furlong W, Weitzman S, Torrance GW. A preference-based approach to health-related quality-of-life for children with cancer. *Internat J Pediatr Hematol/Oncol* 1995;**2**:305–15.

Bjork S. EuroQoL conference proceedings. Swedish Health Economics Institute discussion paper 1, 1991.

Bombardier C, Ware J, Russell I, Larson MG, Chalmers A, Leighton Read J. Auranofin therapy and quality of life in patients with rheumatoid arthritis. *Am J Med* 1986;**81**:565–78.

Bombardier C, Raboud J. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. The Auranofin Cooperating Group. *Control Clin Trials* 1991;**12**:243S–56S.

Boyle MH, Torrance GW. Developing multiattribute health indexes. *Med Care* 1984;**22**:1045–57.

Boyle MH, Torrance GW, Sinclair JC, Horwood SP. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *New Engl J Med* 1983;**308**:1330–7.

Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the Health Utilities Index – Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. *Q Life Res* 1995;**4**:249–57.

Bradlyn AS, Harris CV, Warner JE, Ritchey AK, Zaboy K. An investigation of the validity of the Quality of Well-Being Scale with pediatric oncology patients. *Health Psychol* 1993;**12**:246–50.

Brazier J, Jones N, Kind P. Testing the validity of the EuroQoL and comparing it with the SF-36 health survey questionnaire. *Q Life Res* 1993;**2**:169–80.

Brazier J, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and EuroQoL on an elderly population. *Q Life Res* 1996;**5**:195–204.

Brazier JE, Usherwood TP, Harper R, Thomas K. Deriving a preference based single index from the UK SF-36 Health Survey. *J Clin Epidemiol* 1998;**51**(11):1115–29.

Brooks RG, Jendteg S, Lindgren B, Persson U, Bjork S. EuroQoL: health-related quality of life measurement. Results of the Swedish questionnaire exercise. *Health Policy* 1991;**18**:37–48.

Bryan S, Parkin D, Donaldson C. Chiropody and the QALY – a case-study in assigning categories of disability and distress to patients. *Health Policy* 1191;**18**:169–85.

Bush JW, Anderson JP, Kaplan RM, Blischke WR. Counter-intuitive preferences in health-related quality-of-life measurement. *Med Care* 1982;**20**:516–25.

Cadman D, Goldsmith C. Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. *J Chron Dis* 1986;**39**:643–51.

Cadman D, Goldsmith C, Bashim P. Values, preferences and decisions in the care of children with developmental disabilities. *Dev Behav Paediatr* 1984;**5**:60–4.

Calfas KJ, Kaplan RM, Ingram RE. One-year evaluation of cognitive-behavioural intervention in osteoarthritis. *Arthritis Care Res* 1992;**5**:202–9.

Caperna J, Mathews WC. 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. *J Invest Med* 1996;**44**:A155.

Carr-Hill RA. A good measure for Eurohealth? *Health Serv J* 1991;**101**:24–5.

Carr-Hill RA. A second opinion: health-related quality of life measurement – Euro style. *Health Policy* 1992;**20**:321–8.

Carr-Hill RA, Morris J. Current practice in obtaining the "Q" in QALYs: a cautionary note. *BMJ* 1991;**303**:699–701.

Chan CLH, Villar RN. Obesity and quality-of-life after primary hip-arthroplasty. *J Bone Joint Surg – Br Vol* 1996;**78B**:78–81.

Coast J. Reprocessing data to form QALYs. *BMJ* 1992;**305**:87–90.

Cole RP, Shakespeare V, Shakespeare P, Hobby JA. Measuring outcome in low-priority plastic surgery patients using Quality of Life indices. *Br J Plast Surg* 1994;**47**:117–21.

Currim IS, Sarin RK. A comparative evaluation of multi-attribute consumer preference models. *Management Sci* 1984;**30**(5):543–61.

de Groot J, de Groot W, Kamphuis M, Vos PF, Berend K, Blankestijn PJ. Kwaliteit van leven van dialysepatienten in Utrecht en Willemstad weinig verschillend [Little difference in quality of life of dialysis patients in Utrecht and Willemstad]. *Ned Tijdschr Geneeskd* 1994;**138**:862–6.

Dirksen SR. Search for meaning in long-term cancer survivors. *J Adv Nurs* 1995;**21**:628–33.

Dolan P. Search for a critical-appraisal of EuroQoL – a response by the EuroQoL group to Gafni and Birch. *Health Policy* 1994;**28**:67–9.

Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQoL: results from a UK general population survey. Centre for Health Economics discussion paper 138. York: University of York, 1995.

Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996;**2**:209–32.

Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 1987.

Elvik R. The validity of using health state indexes in measuring the consequences of traffic injury for public-health. *Soc Sci Med* 1995;**40**:1385–98.

Erickson P, Kendall EA, Anderson JP, Kaplan RM. Using composite health status measures to assess the nation's health. *Med Care* 1989;**27**:S66–76.

Essink-Bot ML, Bonsel GJ, Van Der Maas PJ. Valuation of health states by the general public: feasibility of a standardized measurement procedure. *Soc Sci Med* 1990;**31**:1201–6.

Essink-Bot ML, Stouthard ME, Bonsel GJ. Generalizability of valuations on health states collected with the EuroQoLc-questionnaire. *Health Econ* 1993;**2**:237–46.

Essink-Bot ML, Vanroyen L, Krabbe P, Bonsel GJ, Rutten FFH. The impact of migraine on health-status. *Headache* 1995;**35**:200–6.

EuroQoL Group. EuroQoL – a new facility for the measurement of health-related quality-of-life. *Health Policy* 1990;**16**:199–208.

EuroQoL Group. Not a quick fix (response to Carr-Hill). *Health Serv J* 1991;**101**:29.

EuroQoL Group. EuroQoL – a reply and reminder. *Health Policy* 1992;**20**:329–32.

Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. *J Clin Oncol* 1992;**10**:923–8.

Feeny D, Leiper A, Barr RD, Furlong W, Torrance GW, Rosenbaum P, Weitzman S. The comprehensive assessment of health status in survivors of childhood cancer: application to high-risk acute lymphoblastic leukaemia. *Br J Cancer* 1993;**67**:1047–52.

Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. *PharmacoEconomics* 1995;**7**:490–502.

Fryback DG, Dasbach ED, Klein R, Klein BEK, Martin PA, Dorn N, Peterson K. Health assessment by SF-36, Quality of Well-Being Index and time trade-offs: predicting one measure from another. *Med Decis Making* 1992;**12**:348P.

Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K, *et al.* The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making* 1993;**13**:89–102.

Furlong W, Torrance GW, Feeny D. Properties of Health Utilities Index: preliminary evidence. *Qual Life Newslett* 1995;3–10.

Ganiats TG, Palinkas LA, Kaplan RM. Comparison of Quality of Well-Being Scale and Functional Status Index in patients with atrial fibrillation. *Med Care* 1992;**30**:958–64.

Gilbert A, Owen N, Innes JM, Sansom L. Trial of an intervention to reduce chronic benzodiazepine use among residents of aged-care accommodation. *Aust NZ J Med* 1993;**23**:343–7.

Gold M, Franks P, Erickson P. Assessing the health of the nation: the predictive value of a preference based measure and self-rated health. *Med Care* 1996;**34**:163–77.

Gravelle H. Valuations of EuroQoL health states: comments and suggestions. Paper presented at the ESRC/SHHD Workshop on Quality of Life, Edinburgh, unpublished, 1995.

Gudex C. QALYs and their use by the health service. Discussion paper 20. York: Centre for Health Economics, University of York, 1986.

Gudex C, Kind P. The QALY toolkit. Centre for Health Economics discussion paper 93. York: University of York, 1988.

Gudex C, Williams A, Jourdan M, Mason R, Maynard J, O'Flynn R, *et al.* Prioritising waiting lists. *Health Trends* 1990;**22**:103–8.

Gudex C.M. Health-related quality of life in endstage renal failure. *Q Life Res* 1995;**4**:359–66.

Gudex C, Kind P. Chiropody and the QALY – a case-study in assigning categories and distress to patients. *Health Policy* 1991;**19**:79–80.

Gudex C, Kind P, van Dalen H, Durand M-A, Morris J, Williams A. Comparing scaling methods for health state valuations: Rosser revisited. Centre for Health Economics discussion paper 107. York: University of York, 1993.

Holbrook TL, Hoyt DB, Anderson JP, Hollingsworth-Fridlund P, Shackford SR. Functional limitation after major trauma: a more sensitive assessment using the Quality of Well-Being Scale – the trauma recovery pilot project. *J Trauma* 1994;**36**:74–8.

Hollingworth W, Mackenzie R, Todd CJ, Dixon AK. Measuring changes in quality-of-life following magnetic-resonance-imaging of the knee – SF-36, EuroQoL((c)) or Rosser index. *Q Life Res* 1995;**4**:325–34.

Hornberger JC, Redelmeier DA, Petersen J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *J Clin Epidemiol* 1992;**45**:505–12.

Humphreys WV, Evans F, Watkin G, Williams T. Critical limb ischemia in patients over 80 years of age – options in a district general-hospital. *Br J Surg* 1995;**82**:1361–3.

Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H. Validity of EuroQoL – a generic health status instrument – in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. *Br J Rheumatol* 1994;**33**:655–62.

Kallis P, Unsworth White J, Munsch C, Gallivan S, Smith EE, Parker DJ, *et al.* Disability and distress following cardiac surgery in patients over 70 years of age. *Eur J Cardiothorac Surg* 1993;**7**:306–11.

Kaplan RM. Health outcome models for policy analysis. *Health Psychol* 1989;**8**:723–35.

Kaplan RM. Application of a general health policy model in the American health care crisis. *J R Soc Med* 1993a;**86**:277–81.

Kaplan RM. Quality of life assessment for cost/utility studies in cancer. *Cancer Treat Rev* 1993b;**19**(Suppl A): 85–96.

Kaplan RM. Value judgement in the Oregon medicaid experiment. *Med Care* 1994a;**32**:975–88.

Kaplan RM. Using quality-of-life information to set priorities in health-policy. *Social Indicators Res* 1994b;**33**:121–63.

Kaplan RM, Atkins CJ. The well-year of life as a basis for patient decision-making. *Patient Educ Couns* 1989;**13**:281–95.

Kaplan RM, Anderson JP. A general health policy model: update and application. *Health Services Res* 1988;**23**:203–35.

Kaplan RM, Bush JW. Health-related quality of life measurement for evaluation research and policy analysis. *Health Psychol* 1982;**1**:61–80.

Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. *Health Serv Res* 1976;**11**:478–507.

Kaplan RM, Bush JW, Berry CC. Health status index: category rating versus magnitude estimation for measuring levels of well-being. *Med Care* 1979;**17**:501–25.

Kaplan RM, Atkins CJ, Timms R. Validity of a quality of well-being scale as an outcome measure in chronic obstructive pulmonary disease. *J Chronic Dis* 1984;**37**:85–95.

Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F, Orenstein D. The Quality of Well-being Scale. Applications in AIDS, cystic fibrosis, and arthritis. *Med Care* 1989;**27**:S27–43.

Kaplan RM, Anderson JP, Wingard DL. Gender differences in health-related quality of life. *Health Psychol* 1991a;**10**:86–93.

Kaplan RM, Debon M, Anderson BF. Effects of number of rating scale points upon utilities in a Quality of Well-Being scale. *Med Care* 1991b;**29**:1061–4.

Kaplan RM, Coons SJ, Anderson JP. Quality of life and policy analysis in arthritis. *Arthritis Care Res* 1992;**5**:173–83.

Kaplan RM, Anderson JP, Patterson TL, Mccutchan JA, Weinrich JD, Heaton RK, *et al.* Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Centre. *Psychosom Med* 1995;**57**:138–47.

Kind P. Measuring valuations for health states: a survey of patients in general practice. Centre for Health Economics discussion paper 76. York: University of York, 1990.

Kind P. An interim tariff for EuroQoL health states. Personal communication, 1994.

Kind P. The EuroQoL instrument: an index of health – related quality of life. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials, 2nd edn. Philadelphia: Lippincott-Rivera, 1996:191–201.

Kind P, Gudex C. The HMQ: measuring health status in the community. Centre for Health Economics discussion paper 93. York: University of York, 1991.

Kind P, Gudex CM. Measuring health-status in the community – a comparison of methods. *J Epidemiol Commun Health* 1994;**48**:86–91.

Kind P, Rosser R. The quantification of health. *Eur J Social Psych* 1988;**18**:63–77.

Kind P, Gudex C, Dolan P, Williams A. Practical and methodological issues in the development of the EuroQoL: the York experience. In: Albrecht GL, Fitzpatrick R, editors. Advances in medical sociology. Greenwich, 1994:219–253.

Launois R, Henry B, Marty JR, Gersberg M, Lassale C, Benoist M, *et al.* Chemonucleolysis versus surgical diskectomy for sciatica secondary to lumbar disc herniation – a cost and quality-of-life evaluation. *PharmacoEconomics* 1994;**6**:453–63.

Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopaedic evaluation. *Med Care* 1990;**28**:632–42.

Lonnqvist J, Sintonen H, Syvalahti E, Appelberg B, Koskinen T, Mannikko T, *et al.* Antidepressant efficacy and quality of life in depression: a double-blind study with moclobemide and fluoxetine. *Acta Psychiatr Scand* 1994;**89**:363–9.

Lonnqvist J, Sihvo S, Syvalahti E, Sintonen H, Kiviruusu O, Pitkanen H. Moclobemide and fluoxetine in the prevention of relapses following acute treatment of depression. *Acta Psychiatr Scand* 1995;**91**:189–94.

Mackenzie R, Hollingworth W, Dixon AK. Quality of life assessments in the evaluation of magnetic resonance imaging. *Q Life Res* 1994;**3**:29–37.

Magee TR, Scott DJ, Dunkley A, St Johnston J, Campbell WB, Baird RN, *et al.* Quality of life following surgery for abdominal aortic aneurysm. *Br J Surg* 1992;**79**:1014–16.

Manzetti JD, Hoffman LA, Sereika SM, Sciurba FC, Griffith BP. Exercise, education, and quality of life in lung transplant candidates. *J Heart Lung Transplant* 1994;**13**:297–305.

Mold JW, Holtgrave DR, Bisonni RS, Marley DS, Wright RA, Spann SJ. The evaluation and treatment of men with asymptomatic prostate nodules in primary care: a decision analysis. *J Fam Pract* 1992;**34**:561–8.

Measurement and Valuation of Health Group The measurement and valuation of health: first report on the main survey. York: Centre for Health Economics, University of York, 1994.

Measurement and Valuation of Health Group The measurement and valuation of health: Final report on the modelling of valuation tariffs. York: Centre for Health Economics, University of York, 1995.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Plan Manag* 1991a;**6**:234–42.

Nord E. EuroQoL – health-related quality-of-life measurement – valuations of health states by the general public in Norway. *Health Policy* 1991b;**18**:25–36.

Nord E. Unjustified use of the Quality of Well-Being Scale in priority setting in Oregon. *Health Policy* 1993;**24**:45–53.

Normantaylor FH, Palmer CR, Villar RN. Quality-of-life improvement compared after hip and knee replacement. *J Bone Joint Surg – Br Vol* 1996;**78B**:74–7.

O'Hanlon M, Fox Rushby J, Buxton MJ. A qualitative and quantitative comparison of the EuroQoL and time-trade-off techniques. *Int J Health Serv* 1994;**5**:85–97.

Orenstein DM, Nixon PA, Ross EA, Kaplan RM. The quality of well-being in cystic fibrosis. *Chest* 1989;**95**:344–7.

Orenstein DM, Pattishall EN, Nixon PA, Ross EA, Kaplan RM. Quality of well-being before and after antibiotic treatment of pulmonary exacerbation in patients with cystic fibrosis. *Chest* 1990;**98**:1081–4.

Orenstein DM, Kaplan RM. Measuring the quality of well-being in cystic fibrosis and lung transplantation. The importance of the area under the curve. *Chest* 1991;**100**:1016–18.

Parkin D. Valuing health states: an exploratory data analysis approach. Paper presented to a meeting of the Health Economists Study Group, University of Oxford, 1991.

Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res* 1973a;**8**:228–45.

Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. *J Health Soc Behav* 1973b;**14**:6–23.

Payne SP, Galland RB. The use of a simple clinical cardiac risk index predictive of long-term outcome after infrarenal aortic reconstruction. *Eur J Vasc Endovasc Surg* 1995;**9**:138–42.

Petrou S, Davey P, Malek M. The application of the Rosser–Kind classification to hip and knee joint replacement surgery. Health Economists Study Group Paper, 1992.

Rabin R, Rosser RM, Butler C. Impact of diagnosis on utilities assigned to states of illness. *J R Soc Med* 1993;**86**:444–8.

Rawles J, Light J, Watt M. Quality of life in the first 100 days after suspected acute myocardial infarction – a suitable trial endpoint? *J Epidemiol Community Health* 1992;**46**:612–16.

Read JL, Quinn RJ, Hoefer MA. Measuring overall health: an evaluation of three important approaches. *J Chronic Dis* 1987;**40** (Suppl 1):7S–26S.

Reed PG. Religiousness among terminally ill and healthy adults. *Res Nurs Health* 1986;**9**:35–41.

Rissanen P, Aro S, Slatis P, Sintonen H, Paavolainen P. Health and quality of life before and after hip or knee arthroplasty. *J Arthroplasty* 1995;**10**:169–75.

Rissanen P, Aro S, Sintonen H, Slatis P, Paavolainen P. Quality-of-life and functional ability in hip and knee replacements – a prospective-study. *Q Life Res* 1996;**5**:56–64.

Rosser RM, Kind P. A scale of valuations of states of illness: is there a social consensus? *Int J Epidemiol* 1978;**7**:347–58.

Rosser R, Sintonen H. The EuroQoL quality of life project. In: Quality of life assessment: key issues in the 1990s. Lancaster: MTP Press, 1993:197–9.

Rosser RM, Watts VC. The measurement of hospital output. *Int J Epidemiol* 1972;**1**:361–8.

Rosser R, Allison R, Butler C, Cottee M, Rabin R, Selai C. The Index of Health-related Quality of Life. In:Hopkins A, editor. Measures of the quality of life and the uses to which such measures may be put. London: Royal College of Physicians of London, 1992.

Rosser R, Allison R, Butler C, Cottee M, Rabin R, Selai C. The Index of Health-related Quality of Life (IHQL): a new tool for audit and cost-per-QALY analysis. In: Walker SR, Rosser RM, editors. Quality of life assessment: key issues in the 1990s. Lancaster: MTP Press, 1993:179–84.

Saigal S, Feeny D, Furlong W, Rosenbaum P, Burrows E, Torrance G. 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. *J Pediatr* 1994;**125**:418–25.

Saigal S, Rosenbaum PL, Furlong WJ, Feeny DH, Burrows E. Self-assessment of their own health-status by extremely low-birth-weight and control teenagers using a multiattribute health-status classification-system. *Paediat Res* 1995;**37**:A271.

Schneiderman LJ, Kronick R, Kaplan RM, Anderson JP, Langer RD. Effects of offering advance directives on medical treatments and costs. *Ann Intern Med* 1992;**117**:599–606.

Sculpher M, Bryan S, Dwyer N, Hutton J, Stirrat GM. An economic evaluation of transcervical endometrial resection versus abdominal hysterectomy for the treatment of menorrhagia. *Br J Obstet Gynaecol* 1993;**100**:244–52.

Selai C, Rosser R. Eliciting EuroQoL descriptive data and utility scale values from inpatients – a feasibility study. *PharmacoEconomics* 1995;**8**:147–58.

Sintonen H. An approach to measuring and valuing health states. *Soc Sci Med* 1981;**15C**:55–65.

Sintonen H. Terveyteen liittyvan elamanlaadun mittaamisesta [Health-related quality of life measures]. *Sairaanhoitaja* 1993;17–19.

Spiegelhalter DJ. The choice of "tariff": comments on the measurement and valuation of health project. Paper presented at the ESRC/SHHD Workshop on Quality of Life, Edinburgh, unpublished, 1995.

Tandon PK, Stander H, Schwarz RP Jr. Analysis of quality of life data from a randomized, placebo-controlled heart-failure trial. *J Clin Epidemiol* 1989;**42**:955–62.

Thomas R, Thomson K. Health-related quality of life: technical report. London: SCPR, 1992.

Torrance GW, Boyle MH, Horwood SP. Applications of multi-attribute utility theory to measure social preferences for health states. *Operat Res* 1982;**30**:1043–69.

Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health Utilities Index. *PharmacoEconomics* 1995;**7**:503–20.

Tramarin A, Milocchi F, Tolley K, Vaglia A, Marcolini F, Manfrin V, *et al.* An economic evaluation of home-care assistance for AIDS patients: a pilot study in a town in northern Italy. *Aids* 1992;**6**:1377–83.

Unsworthwhite J, Kallis P, Treasure T, Pepper JR. Quality-of-life after cardiac-surgery in patients over 70 years of age. *Cardiol Elderly* 1994;**2**:133–8.

van-Agt HM, Essink-Bot ML, Krabbe PF, Bonsel GJ. Test–retest reliability of health state valuations collected with the EuroQoL questionnaire. *Soc Sci Med* 1994;**39**:1537–44.

van Dalen H, Williams A, Gudex C. Lay peoples evaluations of health – are there variations between different subgroups. *J Epidemiol Commun Health* 1994;**48**:248–53.

Visser MC, Fletcher AE, Parr G, Simpson A, Bulpitt CJ. 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. *J Clin Epidemiol* 1994;**47**:157–63.

Verhoef CG, Verbeek AL, Stalpers LJ, van Daal WA. Utiliteitsmeting bij de klinische besluitvorming [Utility assessment in clinical decision making]. *Ned Tijdschr Geneeskd* 1990;**134**:2195–200.

Wade DT. The Q in QALYs. *BMJ* 1991;**303**:1136–7.

Watkins LD, Bell BA, Marsh HT, Uttley D. A scale for neurosurgical audit. *Br J Neurosurg* 1990;**4**:463–5.

Whynes DK, Neilson AR. Convergent validity of two measures of the quality of life. *Health Econ* 1993;**2**:229–35.

Whynes DK, Neilson AR, Robinson MH, Hardcastle JD. Colorectal cancer screening and quality of life. *Q Life Res* 1994;**3**:191–8.

Williams A. Economics of coronary artery bypass grafting. *BMJ* 1985;**291**:326–9.

Williams A. The measurement and valuation of health: a chronicle. Centre for Health Economics discussion paper 136. York: University of York, 1995a.

Williams A. The role of the EuroQoL instrument in QALY calculations. Centre for Health Economics Discussion paper 130. York: University of York, 1995b.

Wu AW, Mathews WC, Brysk LT, Hampton Atkinson J, Grant I, Abramson I, *et al.* Quality of life in a placebo-controlled trial of Zidovudine in patients with AIDS and AIDS-related complex. *J Acquired Immune Deficiency Syndromes* 1990;**3**:683–90.

Chapter 6

The use of non-preference-based measures of health in economic evaluation

Non-preference-based measures of HRQoL, often referred to in the literature as HSMs, are increasingly used in clinical trials to assess the efficacy and effectiveness of healthcare interventions in terms of patient-perceived health. They provide an important source of data regarding the benefits of health care but were not designed for use in economic evaluation. Some health economists have attempted to use them in conducting economic evaluations alongside clinical trials (e.g. Buxton *et al.*, 1985; Nichol *et al.*, 1992). However, the use of HSMs in economic evaluation has either been criticised by health economists, largely because they do not explicitly incorporate preferences (Culyer, 1978; Williams, 1989; Johannesson *et al.*, 1996) or ignored by them. For reasons reviewed below, we believe HSMs will continue to be widely used in clinical trials and are likely to continue to be far more popular than economic measures of benefit. It is therefore important to examine the potential use of HSMs in economic evaluation in order to extend the scope for undertaking such analyses alongside clinical trials.

This chapter begins by reviewing the characteristics of a sample of HSMs. It then examines why HSMs are used more than economic measures. The economic criticisms of using HSMs in economic evaluation are then reviewed, along with the evidence on the relationship between HSMs and preference-based measures. On the basis of these sections we make recommendations regarding the use of HSMs in economic evaluation. The last section considers whether it is possible to further develop or adapt these measures for use in economic evaluation.

Search strategy and methods of review

The abstracts of the 155 papers identified in chapter 3 have been used for the review of the use of HSMs presented below. As before, the papers have not been systematically reviewed against quality criteria. Nonetheless, it is intended to be a comprehensive review and one which presents an accurate balance of opinion (we have tried to reflect disagreements rather than to hide them) from the economics literature, but it inevitably contains our own judgements and opinions.

The search strategy and methods of a review of papers comparing HSMs with preference measures are presented later in this chapter.

Characteristics of HSMs

The term 'HSM' is used here to describe instruments designed to measure quantitatively dimensions of health thought to be of relevance to patients with health problems, caused either by disease, the treatment of disease or other processes such as natural ageing, trauma and pregnancy. This would exclude biomedical measures (such as blood pressure, FEV or cholesterol levels) or diagnostic instruments. HSMs can be 'generic' and hence designed for use across all conditions or specifically designed for a particular disease. Such measures have been available since the 1940s (Karnofsky and Burchenal, 1949), but did not become widely used until the 1960s and 1970s. By 1987, there were over 200 HSMs identified by Spilker *et al.* (1990).

HSMs vary widely in terms of content, format and scaling. The principal features of a sample of eight HSMs are presented in *Table 3*. The instruments have been selected to demonstrate the diversity of measures in terms of their size, coverage of health domains, method of administration, and sources of values, not for being typical or even representative.

The contents vary considerably between the measures, from generic concepts of functioning through to specific symptoms (e.g. dyspnoea for respiratory disease, dexterity for arthritis and so forth). The methods of completing the questionnaires include clinical interview, professional assessment, researcher interview and self-completion, either in the clinic or at home. Many of these questionnaires are completed by the patient. Though this is not typical, it has become more common in recent years. The

TABLE 3 *Characteristics of five health status measures*

Questionnaire	No. of dimensions	Description of dimensions/items	No. of items	Source of responses	Method of administration	Source of values	Results
Condition specific							
St. George's Respiratory Questionnaire	4	Symptoms (e.g. shortness of breath and wheezing), activity (e.g. walking and playing games), impacts (e.g. embarrassment)	50	Patient	Interview or self-completion	Patients (using VAS)	Profile and index
Chronic Respiratory Questionnaire	4	Dyspnoea, fatigue, emotional function, mastery	20	Patient	Interview	Assumed	Profile
Barthel	1	Mobility, grooming, dressing, continence	10	Professional	Professional assessment	Assumed	Index
Generic							
SF-36	8	Physical functioning, role limitations (physical and emotional problems), social fucntioning, pain, mental health, general health perception	36	Patient or proxy	Self-completion, interviewer administration	Assumed	Profile
NHP	6	Mobility, social isolation, pain, emotional reactions, energy	38	Patient	Self-completion	Thurstone's method	Profile

Chronic Respiratory Questionnaire incorporates a further development, where patients are asked to identify the important activities which make them breathless, as well as providing the assessment. The developers have argued that this approach has the advantage of generating a score more responsive to health change (Guyatt *et al.*, 1993), though it is of doubtful use in interpersonal comparisons.

Item responses typically have a simple numeric scaling, such as from 1 to 5, and these scores are summed across the items to derive scores for each dimension (e.g. the SF-36 or the Chronic Respiratory Questionnaire) and/or across all items to derive an overall score (e.g. the Barthel measure). This procedure has been mistakenly described in the psychometric literature as being 'unweighted' (Jenkinson, 1991), yet it implicitly assumes **equal** weighting. In others, such as the St. George's Respiratory Questionnaire and the SIP, weights have been derived by explicit valuation procedures.

We do not include any of the 'QALY' instruments in this list, such as the EQ-5D, since these are purporting to value health rather than simply to measure it. Nonetheless, some of these measures do share many of the characteristics found in HSMs, including the dimensions and items, and the methods of administration. These QALY measures were reviewed in chapter 4.

Why consider the use of HSMs 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 healthcare interventions, for example, has been limited (Backhouse *et al.*, 1992) and certainly not sufficient to provide a complete and up-to-date assessment of the cost-effectiveness of health technologies (Drummond *et al.*, 1993), whereas the use of condition-specific, and to a lesser extent generic, HSMs have become more widespread. This is largely because the trials were designed to address clinical rather than economic questions. Yet even amongst researchers who are seeking to address a broader set of questions, including 'cost-effectiveness', there has been a reluctance to use economic measures. This reluctance may in part be the consequence of continued unfamiliarity with economic measures.

Drummond and Davies (1991) have identified three explanations for this reluctance to use such measures amongst clinical researchers. The first is the additional burden from using any extra measures in clinical trials which increase costs and risk burdening the patient. In defence of preference-based measures, many of the instruments used by economists take less time than many HSMs (e.g. the EQ-5D takes less than

3 minutes to complete), and cost considerably less than many clinical tests. A researcher intending to conduct an economic evaluation may have to consider reducing the burden of clinical measures. A second concern with using preference-based measures is that preference elicitation techniques, such as SG and TTO, may be distressing to patients. It may result in patients withdrawing from the trial and is questionable ethically. This difficulty would be avoided by using one of the QALY classifications. The third and final concern, which is potentially a more fundamental problem, is the view that preference-based measures suffer from being insensitive or even irrelevant for many conditions. Condition-specific HSMs are argued to contain more relevant health dimensions and hence be more responsive to changes in health in patients with the condition (Guyatt *et al.*, 1987), while the generic HSMs tend to be larger than the QALY HSCs and hence have more scope to measure change. This is a concern shared by some health economists (Donaldson *et al.*, 1988; Hall *et al.*, 1992).

Care must be taken in reviewing the claims for the greater sensitivity of HSMs over preference-based measures since they are often based on the psychometric criteria of construct validity and responsiveness. As discussed in chapter 3, these are not appropriate criteria for testing the validity of a measure for use in economic evaluation, since they take no account of the importance of any differences in health. An HSM may be found to have a larger effect size for a given health change, but this does not mean it is a better reflection of preferences. Nonetheless, relevance and sensitivity are important components of the descriptive validity of a measure, and there is evidence to support the claim that QALY HSCs and direct utility assessment can be insensitive to important health changes in some patient groups (see chapter 4).

In summary, HSMs are far more widely used than preference-based measures in health services research for a number of reasons. It is therefore extremely pertinent to ask the question as to whether such measures can be used to assess the relative efficiency of healthcare interventions.

Economic criticisms of non-preference-based health measures

Assessing the validity of HSMs for use in economic evaluation is concerned with establishing the extent to which they reflect preferences. The goal of the developers of HSMs has been to **measure** various concepts of health. The use of HSMs in economic evaluation depends on the extent to which HSM scores reflect the intensity of peoples' preferences for health changes from healthcare interventions. The distinction between the aims of measuring or numerically describing health and the estimation of peoples' **preferences** for health is essential to understanding the economic critique of HSMs. We begin this section by reviewing HSMs in terms of the validity of the descriptions and then consider the issue of values.

Descriptive validity: choice of dimensions and items

The choice of dimensions and items is an important value judgement. The exclusion of a dimension is equivalent to assigning it a value of 0. This may not matter if this is indeed found to be the case or the dimension is unaltered by the healthcare intervention being evaluated. However, it is rare for either of these to be demonstrated.

The methods of selecting items and dimensions include using expert opinion (i.e. the designer and/or a panel of experts), reviewing the literature (including existing measures, such as done for the SF-36), eliciting patient views (as ascertained in interviews and surveys), and statistical methods (e.g. factor analysis). The most common method is the use of expert opinion. Statistical methods of item selection take account of the internal consistency or homogeneity of items within dimension, including the use of factor analysis to identify clusters of related items. These statistical approaches are based on the correlation of items and hence may have little relationship to preferences and can lead to the exclusion of important items simply because they did not fit neatly into the hypothesised domains (see chapter 3). Economists concerned with correctly reflecting the individual's preferences are likely to prefer a patient based approach to generating and selecting dimensions and their items.

These comments are not intended to suggest that HSMs have been worse than preference-based measures in the methods of selecting dimensions. Indeed the methods employed have often been more thorough.

Valuation

Scoring of HSMs

For HSMs there can be three components to the scoring: (1) scores are assigned to the response choices offered in each question (e.g. the SF-36

physical functioning dimension items have three responses – 'limited a lot', 'limited a little' and 'not limited at all' – and these are coded 1, 2, and 3, respectively; (2) weightings are used to combine the items to derive a dimension score; and (3) dimensions are combined into an overall total score using a set of weights (though this is not done for many HSMs). The most common method of scoring is to assume equal intervals for each of these components to the scoring (e.g. the Chronic Respiratory Questionnaire, Barthel measure, AIMS and SF-36). Dimension scores are computed by giving equal weight to each item, and for those HSMs which generate a single index, the dimension scores are combined assuming equal weighting.

The arbitrary nature of the assumptions underlying each stage of the scoring has long troubled economists (Culyer, 1978; Torrance, 1986). 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 from the SF-36, the intervals for an item on how much bodily pain a person has had in the last 4 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 in a trial, a reduction in pain from 'mild' to 'very mild' would be equivalent to a reduction from 'severe' to 'moderate'. Yet recent evidence using VAS and SG valuation techniques suggests that patients are unable to perceive a significant difference between 'very mild' and 'mild' but that there is a very large and significant difference between 'moderate' and 'severe' (Brazier *et al.*, 1996). The summing of item scores makes equally untenable assumptions. In the physical functioning scale of the SF-36 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. Given the lack of any empirical basis for these assumptions there must be doubts about even the ordinal properties of these scales as indicators of peoples' preferences, particularly over small changes in the dimension scores. Williams (1989) has gone so far as to suggest that the use of arbitrary weights in some HSMs 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'.

Developers of some HSMs have estimated weights for the items of their instruments, but none have a basis in economic theory and would not perform well against the check-list presented in the previous chapter. For example, the St. George's Respiratory Questionnaire and the SIP have weights derived from asking groups of patients to value the importance of each item on a VAS. Nonetheless, these are more likely to possess ordinal properties.

For many clinical purposes, it is useful to present separate scores by dimension. To undertake economic evaluation, however, it will often be necessary to be able to combine the dimensions into an overall indicator of preferences. However, the generation of a single index score for health has been opposed by many developers of HSMs. The developers of the NHP, for example, have argued: 'The simple addition of affirmative responses gives misleading results because of the features of pain, social life, emotion, and so on are qualitatively distinct and made up of different facets which can not have common denominators' (Hunt *et al.*, 1986). This view is understandable when the purpose is to derive a measure of health, but this is not the purpose for use in economic evaluation. A profile measure might indicate an improvement in the physical functioning and possibly other related dimensions such as social functioning but a deterioration in the pain. At the end of a clinical trial it would not be possible to determine whether the treatment was effective, let alone whether it was cost-effective. A trade-off needs to be made between dimensions in order to decide whether the patient should have the treatment and this is not possible with profile HSMs.

Some HSMs do combine the different dimensions to form a single index (e.g. the St. George's Respiratory questionnaire, the SIP and the Barthel measure). As for the aggregation of items, many assume an equal weighting between dimensions (e.g. the Barthel measure), while others combine the items using item weights estimated using valuation techniques such as the VAS. In addition to the previous criticisms of these methods, the scoring systems make an assumption of simple additivity between dimensions, where the value of one dimension is assumed to be unaltered by the level of another dimension (Culyer, 1978). This rules out the prospect of any interaction between 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, an interaction may increase the deficit over and above the sum of the two parts.

The equal interval assumptions underlying most HSMs have been defended by psychometricians and other health services researchers. The relative importance of the different health concepts is in part taken account of by the number of items used to represent them. Thus in the case of the physical functioning dimension of the SF-36, there are three items for walking the block against one for going shopping. Others have argued that there is no theoretical or empirical basis for using anything other than unit weightings, and so equal weighting is favoured on the grounds of a default (Fletcher *et al.*, 1992). It has also been claimed that it makes little difference in practice whether or not equal interval weighting is used in the case of a widely used generic HSM, the NHP (Jenkinson, 1991). To economists, there are convincing theoretical reasons for supposing weightings are not equal, but ultimately these arguments are open to empirical testing.

Time and risk

The outcome of a treatment is often estimated as the mean difference between health scores before and after treatment of patients in the trial. A more sophisticated approach to analysing repeated measures is to estimate the health change as the difference between the mean pretreatment scores and a weighted average of mean scores across the post-treatment assessments, with the weights proportional to the time between each assessment (Matthews *et al.*, 1996). This method of analysis ignores the impact of both time and risk on peoples' preferences for different outcomes (O'Brien, 1994).

Time can have implications for the value of a health state. The conventional method for analysing repeated measures assumes independence of the duration spent in a state, when the state occurs, and which states precede or follow it. Criticisms of these assumptions have been made of QALYs and are summarised in chapter 2.

The analysis of time raises another problem. An important limitation of HSM scores is that they do not include mortality (Feeny *et al.*, 1990) This can lead to a statistical artefact whereby an improvement in survival can bring down the mean health status of the cohort simply because the survivors have a worse state of health. For economic evaluation and clinical decision-making it is often necessary to combine survival with health status, such as for benign prostatic hyperplasia, where the surgical intervention is associated with a risk of fatality, or in the treatment of terminal conditions where a treatment to extend life is associated with unpleasant side-effects; only QALYs and HYEs do this.

There can be a very wide range of outcomes for common treatments such as cholecystectomy, with major negative health effects from complications (Nicholl *et al.*, 1992) and mortality. Conventional analysis of HSM data assumes people are risk-neutral. Yet in health care there is evidence that many people are averse to risk (Loomes and McKenzie, 1989). Patients may choose a treatment which achieves a lower expected or mean improvement in the HSM scores than another, but is associated with less variance. The distribution of health outcomes should not be ignored when comparing the effectiveness of treatments.

A review of empirical comparisons of preference- and non-preference-based measures

In the previous section we reviewed the reasons why HSMs are unlikely to reflect preferences. Whether or not they do is ultimately an empirical question. There are studies in which HSMs are used in studies alongside preference-based measures, and this presents us with the opportunity to examine the empirical relationship between them. Whilst there are many examples of studies using these measures in combination, relatively few studies have attempted to explore or explain this relationship. In this next section we review work done to date which has investigated the relationship between HSMs and utility-based measures of HRQoL.

Revicki and Kaplan's review

The relationship between HSMs and preference measures has been explored in a number of studies, some of which are clinical trials and cover a number of diseases. Other papers are methodological investigations of different outcome measures. A review of these papers has already been published by Revicki and Kaplan (1993). The authors of the review conducted a MEDLINE search to identify studies which included both psychometric and utility measures **and** which also gave correlations between the two types of measure. This search identified 15 studies published between January 1985 and March 1993.

The studies identified employed the QWB (47%) and TTO (40%) approaches most commonly,

while SG was only used in 2/15 studies. The HSMs used included the SIP, SF-36, General Health Rating Index, Spitzer Quality of Life Index, Karnofsky Scale and the Specific Activity Scale. Correlations between measures were summarised by the preference measurement method. Firstly for the RS or VAS, preferences and HSM scores were correlated 0.17–0.46, with 3–21% of the variance in the VAS score being predicted by individual HSM scores. It was estimated by regression analysis that 27–34% of the VAS values could be accounted for by a combination of health status scores.

For the TTO method (five studies), it was estim-ated that the correlation between HSM scores and TTO preferences was between 1 and 43%. In most studies there was only small to moderate correlation between TTO and various HSM measures. For the SG method (three studies) utilities were poorly to moderately correlated (r = 0.01–0.3) to HSM measures. Approximately 1–25% of variance according to which HSM was used was shared between HSMs and SG.

There were six studies which used the QWB as the preference measure. It was found that 11–50% of variance was shared between HSMs and QWB scores. QWB scores were correlated more with measures of physical as opposed to psychological functioning. Only one study (Brazier *et al.*, 1993) compared the dimensions of a multi-attribute preference-based measure (EQ-6D) and an HSM (SF-36). Spearman rank correlations between the EQ and SF-36 were between 0.48 and 0.60.

In summary, Revicki and Kaplan concluded that there was only a low to moderate association between HSM measures and preference-based measures. The VAS method was more closely correlated with various HSMs than any of the other methods compared in the review, but even here correlations were not impressive, and, as already discussed, there are major doubts about whether the VAS technique can be regarded as a measure of preferences. The correlation of HSMs with SG was particularly poor.

Update

An almost infinite number of permutations of terms from these two concepts meant that it was more feasible to search only for preference-based measure terms. Retrieved papers were then assessed by the expert reviewers to determine if they also contained a health status questionnaire concept.

A recognised search technique for situations where indexing inadequately reflects search concepts is known as 'citation pearl growing' (Hartley *et al.*, 1990). This is when a relevant article is retrieved and then the title and abstract reviewed for the occurrence of free text expressions which are subsequently added to a search strategy. Given that the intention, in the case of this review, is to identify a body of literature an expanded variation of this technique was used, 'citation cluster growing' whereby successive relevant terms are entered into the search strategy until all alternatives are exhausted. Examples of terms identified in this way are given *Box 5*.

Particular challenges for this search strategy resulted from the numerous variants in spelling, terminology and hyphenation as reflected in both the terminology used by authors and the practices used in data entry to the various databases.

This search found four papers written since 1993 (and hence results which are not included in Revicki and Kaplan's paper) that have estimated correlations between HSMs and preference measures:

(1) Revicki *et al.* (1995) compared HSMs and preference measures in patients infected with HIV. The psychometric measures used were an

BOX 5 Search strategy

Health state preference*
Preference based
Preference measure*
Preference weighted measure*
Time preference*
Patient preference*
Standard gamble
Standard reference gamble*
Monte Carlo
Categorical rating {method*/procedure*/scal*}
Category scal*
Categorical scal*
Health state utilit*
Health utilit*
Utility measure*
Utility assessment
Patient utilit*
Time Trade Off
Time Tradeoff
TTO
Contingent Valuation
Willingness to pay
Economic-value-of-life
Discount*

adapted version of the Medical Outcome Study – HIV (MOS-HIV) instrument, The Centre for Epidemiologic Studies (CES) Depression Scale, and the SIP Home Management Scale. Health state utilities were obtained by use of the VAS and SG. The correlations between VAS scores and the MOS-HIV and other health status scores were 0.34–0.56. The highest correlation was between the VAS and the CES Depression Scale of 0.56. The authors found no significant correlation between SG utilities and any of the scores of the HSM.

(2) Rutten-van Molken *et al.* (1995) presented results from a comparison of four instruments in the evaluation of two drug therapies in asthmatics. One of the main aims was to test the construct validity of four measures: the Asthma Quality of Life Questionnaire (AQLQ); the Living with Asthma Questionnaire (LWAQ); the SIP; and SG and VAS ratings. The paper reports Spearman rank correlation coefficients for both VAS and SG utilities against the other three HRQoL instruments. The correlation coefficients for VAS ratings against the other measures were 0.47 for the AQLQ, –0.43 for the LWAQ and –0.59 for the SIP. For SG utilities the corresponding figures were 0.19, –0.13 and –0.15. As in other studies there seems to be a better correlation between VAS values and HSMs than is found between SG values and HSM results.

(3) Tsevat *et al.* (1996) in a study of patients infected with HIV examined the relationship between measures at two points in time. The preference-based measures were TTO, the VAS and the QWB, and health status was measured using the 18-item Mental Health Inventory, the Dyspnea-Fatigue Index (DFI) and the SF-36. For TTO the strongest correlates (0.51–0.59) were with measures of physical functioning (the SF-36 physical functioning score, SF-36 role limitation score, SF-36 vitality and the DFI). For the VAS rating measures, the strongest correlations with HSMs varied between 0.51 (SF-36 physical functioning) and 0.66 (SF-36 general health). A multivariate analysis showed that the SF-36 general health scale, the mental health depression subscale and the DFI accounted for 52% of the variance. The QWB was most strongly correlated with the SF-36 physical functioning (0.51), the DFI (0.67) 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.

(4) Bosch and Hunink (1996) looked at the relationship in patients with intermittent claudication (mild peripheral arterial disease). The HSM was the SF-36, and health preferences were measured using SG, TTO, the VAS and the McMaster HUI. Correlation coefficients between TTO values and the SF-36 ranged 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 HUI and VAS values were more strongly correlated, varying between 0.37 and 0.67. For the HUI, coefficients ranged from 0.40 to 0.60. All dimensions of the SF-36 were significantly correlated with the VAS and HUI values. Regression analyses showed the best combination of SF-36 dimensions explained 28% of variation in TTO scores; the corresponding figures were 14% for the SG technique, 28% for TTO, 53% for the HUI and 61% for the VAS. Overall the relationships between the SF-36 and the TTO and SG techniques were 'poor to moderate', with the correlations between the HUI and the VAS and the SF-36 being described as 'moderate to good'. The authors of this paper compare their findings with other studies which have investigated the relationship between the two types of measure.

Discussion

Overall it would appear that there are only low-to-moderate correlations between HSMs and preference measures, but this was not consistent between or within methods. Most studies find that the relationship is best for the VAS, and VAS-based measure of the QWB. It was argued earlier that the VAS is not regarded as a measure of preferences, but more a measure of health; it is therefore not surprising that VAS ratings are better correlated to measures of health than SG and TTO scores. For TTO and SG, there was a considerable range of values in the size of the correlation, and worst for SG.

The focus in the literature on the relationship in terms of correlations could be criticised since it is regarded as a poor measure of agreement (Bland and Altman, 1988). A high correlation can disguise a poor level of agreement, with the scores of one measure being consistently better or worse than the other. Furthermore, a product moment coefficient assumes a linear relationship exists and this may not be the case. Cairns *et al.* (1991) have explored the potential for establishing an 'exchange rate' between existing condition-specific outcome measures in order to facilitate cross-programme comparisons. A sample of scenarios describing hypothetical patients was

selected from HSMs, and a group of raters was asked to rank them and assign an index number to each using a VAS. They found that the differences in the scores generated by these VASs, however, were not constant between the intervals along the original scales of the three HSMs. Evidence that there is not a simple proportional relationship between the three condition-specific measures was examined. Despite these reservations with the use of correlation coefficients, the basic finding remains of a poor relationship between HSMs and the preference measures of TTO and SG.

Tsevat *et al.* (1996) have argued that 'The existence of such a poor relationship suggests that how people value their health does not correlate with how 'healthy' they are'. There may be different explanations of the poor relationship between HSMs and utility measures depending as to whether we are looking values at '**one point in time**' or comparisons of values measuring change in health **over** time. Tsevat *et al.* (1996) suggest that in the first case the weak relationship is suggestive of the fact that there are 'unmeasured determinants of health **values**'. In the second case, a weak relationship between change in health values and status, it is argued that patients may have adapted to their new health state or have redefined in their minds what constitutes 'excellent health'. Revicki and Kaplan (1993) also offer a number of suggestions as to why such a poor correlation should exist. The process of assigning utility values to health states has a number of features which help to explain the divergence of outcome scores found by using such methods and those found with psychometric approaches. These factors include framing effects (the way in which a scenario is presented), duration of time in the various health states (i.e. time preferences), the inclusion of risk (i.e. treatment outcomes are presented as probabilities in SG), and the general cognitive complexity faced by individuals asked to assign values to health states. QALY scores may also include beliefs about health, emotional factors and indeed 'non-health related factors' such as an individuals personal wealth.

The above analysis leads Revicki and Kaplan to suggest that the two kinds of measure are designed for different purposes and that they are inevitably not interchangeable. Along with others they concluded that the various types of measures are best used alongside each other and that 'A greater understanding of the relationship between preference and health status measures is needed'.

Using non-preference-based health measures in economic evaluation

The previous two sections of this chapter have set out the economic criticisms of HSMs and why HSM scores cannot be regarded as good proxies for preferences. However, HSMs are being widely used in trials, and studies are being published which present costs alongside HSM results (see chapter 7). It is therefore important to examine the use to which HSM scores can be put in economic evaluation.

The usefulness of HSMs in assessing the relative efficiency of interventions depends on the results of the study. In *Table 4* we present seven scenarios

TABLE 4 *Assessing the relative cost-effectiveness of two interventions given different cost and outcome scenarios*

Scenario	Cost	HSM scores	Can cost-effectiveness be evaluated?
1	Lower	Better in at least one dimension and no worse on any other	Yes, by dominance[a]
2	Same	Better in at least one dimension and no worse on any other	Yes[a]
3	Lower	Same across all dimensions	Yes, by cost-minimisation[a]
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

[a] *Given the provisos about the ordinality of the scales*

of costs and outcomes in a comparison of two interventions, and consider whether it is possible to assess their relative cost-effectiveness. 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 it is also straightforward to assess cost-effectiveness, since it is simply a question of choosing the treatment with the better HSM scores since the two have been found to cost the same. The third scenario is the same across all dimensions of the HSM and hence it is a **CMA**. Even for these three scenarios it is necessary to demonstrate the ordinality of the scale of the HSM scores in relation to preferences. The theoretical reasons for doubting that HSMs possess this property were reviewed in an earlier section. The empirical evidence suggests HSMs are significantly if poorly correlated to preference-based measures. This suggests that they should rank states in the same order as preference-based measures, provided there is no trade-off to be made between dimensions. This would indicate that HSM can be useful in assessing cost-effectiveness under each of these three scenarios.

The result is less straightforward for scenarios 4–7, where the usual technique for assessing relative efficiency would be **CEA** where the treatments are compared in terms of their incremental cost-effectiveness ratio. The convention in CEA has been to measure health effects in natural units (Drummond *et al.*, 1987). 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'. However, they 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 to estimate incremental cost-effectiveness ratios across **all** dimensions of the HSM, otherwise it will be necessary to undertake trade-offs between dimension scores which are beyond the scope of these measures. In scenarios 4 and 5 one treatment performs better on same dimensions but worse on others, and hence one treatment could be more of cost-effective on some dimensions but worse on others. Even where one treatment is more cost-effective across all the dimensions of a profile measure, care must be taken in the interpretation. Our review of the evidence found that HSMs do not possess the interval properties required to undertake such comparisons. Furthermore, it is the incremental cost-effectiveness ratio which is important for resource allocation purposes. Therefore, where the least cost-effective intervention costs more and yields a higher benefit, then the greater benefit might be worth the extra cost.

Where there are multiple outcomes, the recommended approach is to present the costs and benefits of the alternatives in a dissaggregated form in a CCA (see chapter 2 for the explanation). This type of presentation might not be helpful because HSM scores have no obvious intuitive meaning. As the developers of the SF-36, for example, 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 HSM scores be compared to other outcomes (such as survival) or cost. Non-preference-based HSMs can not be used to assess cost-effectiveness of the interventions in such circumstances.

This section has described the limited circumstances where HSMs may have a role in assessing relative efficiency. The usefulness of HSMs in economic evaluation depends on the results of the study, but it is usually not possible to predict the results of a study and therefore the advice to researchers designing an economic evaluation is to use preference-based measures alongside the HSMs. An alternative strategy in the longer term would be to adapt HSMs by incorporating preferences and this is considered in the next section.

Developing non-preference-based health measures for use in economic evaluation

The issue addressed here is whether it is possible to develop HSMs in order to utilise their potentially rich source of descriptive information in an economic evaluation. There are five methods for doing this: use arbitrary weights, map from an HSM on to the classification of a preference-based classification, develop exchange rates between HSM scales and preference-based measures, value the items of the HSM using preference-based methods, or use the descriptive data to derive health scenarios for valuation. These will now be reviewed.

Arbitrary weights

One approach is simply to combine the dimension scores or item responses into a single index using an assumed set of weights. A wide range of aggregation schemes could be applied to HSMs, involving the summing of dimension scores or items responses, using different assumed weights. The easiest method would be to weight the dimension scores as follows:

$$\text{health index} = K_1X_1 + K_2X_2 + ... + K_nX_n \qquad (2)$$

where K_j is the dimension weight applied to dimension j, n is the number of dimensions, X_j is the dimension score of dimension j and

$$\sum_{j=1}^{n} K_j = 1$$

An example of this is the work of a team at Brunel University who aggregated the 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: (1) the proportion of affirmative responses to the 38 statements in the NHP; (2) weighting the affirmative responses by weights estimated by the NHP developers, using Thurstone's method of paired comparisons (Hunt *et al.*, 1986); and (3) using unitary statement weights within dimensions and then weighting the dimensions by their proportion of the 38 statements. Similar results were obtained with each method of aggregation, although the range of values examined was very limited and other weighting schemes may have led to different results. Two of the devisors of the NHP, who originally argued against deriving a single index from their profile measure, have recently published a method for obtaining an index of distress for use in conjunction with a measure of dependency in cost–utility studies (McKenna *et al.*, 1993). Their index contains 23 out of the original 38 statements in the NHP (since it excludes mobility) but otherwise is the same as Brunel's first aggregation scheme.

Such arbitrary weighting schemes could easily be applied to HSMs, but they would not generate an index that could be legitimately used in an **economic** evaluation, because the dimension scores are not measures of utility and have not been based on people's preferences. Furthermore, there is no allowance for any possible interaction between the dimensions. Finally, for use in CUA, the index would have to be combined with survival, something O'Brien and colleagues did not feel able to achieve with NHP. The Brunel team 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'.

Mapping on to MAUSs

Another possible way of using non-preference-based HSMs in economic evaluation is to translate responses to the HSM questionnaire into the classification of one of the MAUSs. For a valid translation process to be possible, the non-preference-based HSM must include the dimensions of the MAUS (though it may have more) and to have items which readily equate to the dimension levels of the MAUS. The later would require the HSM items to refer to the same activities and/or severity of a given health problem. The process can be based on the judgements of professionals or researchers, or an explicit set of decision rules can be developed. The validity of any translation procedure could be tested by administering the HSM and the MAUS on the same patients and examining the extent to which the procedure was able to correctly predict the patients position on the MAUS classification.

An interesting attempt at mapping health state descriptors from HSMs on to MAUSs was undertaken by Gudex (1986). She mapped groups of scores from the Ruesch Social Disability Rating Scale (RDR) on to the Rosser matrix, thereby allowing outcome data for patients receiving maintenance haemodialysis to be converted into QALYs. For example, the 'social modifiers' (SMs) dimension of the RDR was converted into the Rosser distress category using the following decision rule; an SM 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) and even that which was produced was of questionable value (Coast, 1992). Furthermore, there has been no attempt to test the predictive accuracy of the translation procedure.

Estimating exchange rates

This would entail estimating a relationship between an HSM and a preference-based measure in order to be able to use HSM scores to predict

preferences. There has been very little work attempting to do this and hence it is not possible to come to any general conclusions. However, the comparisons of HSM scores with preference-based measures suggests that this is not likely to be a promising avenue for development. The only attempt found by this review was the study undertaken by Cairns *et al.* (1991). The differences in the scores generated by the VAS, however, were not constant between the intervals along the original scales of the three HSMs. There is therefore not a simple proportional relationship between the three condition-specific measures examined. 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. Given the theoretical reasons for the differences and the poor correlation between HSMs and preference measures, this is unlikely to be a productive research strategy.

Valuing HSMs

A more radical solution is to completely revalue the content of the HSMs using a preference elicitation technique.

There are different measurement strategies for performing such a task (Froberg and Kane, 1989). One is the holistic approach, whereby all health states defined by the HSM are valued directly. For the smaller instruments, such as the Rosser disability/distress matrix, this is a feasible task since it forms just 29 health states. Most HSMs would be too large for respondents to value all possible states. The review of preference-based measures in chapter 4 identified a number of solutions to this problem. For the QWB and EQ-5D, each respondent valued only a sample of health states defined by the their classification (Kaplan and Bush, 1982). Statistical methods were used to estimate weights for the items. The HUI decomposed the classification and asked respondents to value the single-dimension scales first, and then a sample of health states were used to estimate dimension weights by using MAUT.

The problems with applying these solutions include the fact that many HSMs are far larger than existing HSMs. Furthermore, they were not designed with such a valuation task in mind. The response choices and items of an HSM have no obvious ordinal relationship within dimensions. As a result, for example, the 35 multilevel items of the SF-36 must all be valued and define a total of 2592×1019 unique health states. (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).

The task of estimating a function for such a large and complex classification would be beyond the ability of methods described here. As developers of the SF-36 have commented, '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, 1993). While smaller HSMs, such as the Barthel measure, would not present such a problem, it is likely to be a major undertaking. One solution would be to use only a part of the content of an HSM and thereby substantially simplify the task. This has been attempted by a team in Sheffield who have developed an HSC out of the SF-36 by using just 14 of the 35 items of this instrument, and combining some of the response choices (Brazier *et al.*, 1998). The result is an HSC with six dimensions containing between two and six ranked items defining 9000 states. The disadvantage with this method is that a substantial proportion of information is lost. The implications for the sensitivity of the instrument are not known.

The valuation of HSMs using preference elicitation techniques is feasible, but given the limited research resources it is not likely to be undertaken on most instruments. It is an option for the most widely used generic and possibly condition-specific measures. The latter would permit comparisons within condition but there might be reservations about using them to make cross programme comparisons.

Scenarios

Another approach would be to construct special health state vignettes or whole health scenarios for the outcome of a treatment from the descriptive HSM data. This approach has the advantage of being able to focus on those aspects of health most relevant to the treatment being evaluated. It is also possible to incorporate a time profile of health and hence elicit HYEs (see chapter 2). The methodology for constructing the vignettes from the potentially large volume of descriptive data is not well established. Vignettes have usually been constructed by informal methods using expert opinion, or at best based loosely on qualitative interviews, rather than the evidence procured in a large trial (e.g. Cook *et al.*, 1994). This approach requires further development.

Conclusions

There are many different types of HSMs of variable quality and hence it is difficult to generalise about their role in economic evaluation, but they share the common limitation that they have not been designed for such use. The scoring algorithms usually assume equal intervals between response choices, items and dimensions, and hence are unlikely to reflect preferences. Those HSMs using more complex methods of weighting have not used methods which are likely to reflect preferences on an interval scale. Many generate profiles of dimension scores and therefore there are problems in interpretation when scores change in different directions.

HSMs do have a role in economic evaluation, although it is a limited one. The relative efficiency of interventions can be compared in situations of clear dominance (where cost and outcomes are superior), where interventions cost the same and outcomes are better on one dimension but no worse on any other, and where outcomes are found to be identical and a cost-minimisation can be performed. The absence of interval properties makes it impossible to undertake CEA when trade-offs must be made and hence assess relative efficiency in any other situation. In these circumstances non-preference-based HSMs might be useful in a CCA. For these reasons we recommend that a preference-based measure be used alongside an HSM in trials where it is the intention to undertake an economic evaluation.

The scope for developing HSMs for use in economic evaluation is limited. Crude aggregation is invalid, and mapping HSMs on to QALY instruments is fraught with problems. The derivation of exchange rates between HSM scales and preferences does not look promising since there does not appear to be any simple relationship. The only way forward is to revalue HSMs using preference elicitation techniques. This is a complex and expensive task, and is only likely to be done for the more popular generic instruments. However, this does seem to be an avenue worth pursuing while investigators continue to use HSMs rather than preference-based measures.

References

Backhouse M, Backhouse R, Edey SA. Economic evaluation bibliography. *Health Econ* 1992;**1**(suppl).

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.

Brazier JE. The SF-36 Health Survey and its use in pharmaco-economic evaluation. *PharmacoEconomics* 1995;**7**(5):403–15.

Brazier J, Jones N, Kind P. Testing the validity of the EuroQoL and comparing it with the SF-36 health survey questionnaire. *Q Life Res* 1993;**2**:169–80.

Brazier J, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and EuroQoL on an elderly population. *Q Life Res* 1996;**5**:195–204.

Brazier JE, Usherwood TP, Harper R, Thomas K. Deriving a preference based single index from the UK SF-36 Health Survey. *J Clin Epidemiol* 1998;**51**(11):1115–29.

Bosch JL, Hunink MGM. The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication. *Med Decis Making* 1996;**16**:217–25

Buxton M, Acheson R, Caine N, Gibson S, O'Brien B. 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, 1985.

Cairns J, Johnston K, McKenzie L. Developing QALYs from condition-specific outcome measures. HERU discussion paper 14/91. Aberdeen: University of Aberdeen.

Coast J. Reprocessing data to form QALYs. *BMJ* 1992;**305**:87–90.

Cook J, Richardson J, Street A. A cost–utility analysis of treatment options for gallstone disease – methodological issues and results. *Health Econ* 1994;**3**:157–68.

Culyer AJ. Measuring health: lessons for Ontario. Toronto: University of Toronto Press, 1978.

Drummond MF, Davies L. Economic analysis alongside clinical trials: revisiting the methodological issues. *Int J Tech Assess Health Care* 1991;**7**(4):561–73.

Drummond MF, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 1987.

Drummond M, Torrance G, Mason J. Cost-effectiveness league tables: more harm than good? *Soc Sci Med* 1993;**37**:33–40.

Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: design, analysis, and interpretation. *BMJ* 1992;**305**:1145–8.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – I: measurement strategies. *J Clin Epidemiol* 1989;**42**:345–54.

Gudex C. QALYs and their use by the health service. Discussion paper 20. York: Centre for Health Economics, University of York, 1986.

Guyatt GH, Berman LB, Townend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. *Thorax* 1987;**42**(10):773–8.

Guyatt G, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Int Med* 1993;**118**:622–9.

Hartley RJ, Keen EM, Large JA, Tedd LA. Online searching: principles and practice. London: Bowker Saur, 1990.

Haes JCJM. The integration of quality-of-life and survival – quality-adjusted life years. *Q Life Res* 1993;**2**:60.

Hunt SM, McEwen J, McKenna SP. Measuring health status. London: Croom Helm, 1986.

Jenkinson C. Why are we weighting? A critical examination of the use of item weights in a Health Status Measure. *Soc Sci Med* 1991;**27**:1413–16.

Johannesson M, Jonsson B, Karlsson G. Outcome measurement in economic evaluation *Health Economics* 1996;**5**(4):279–96.

Kaplan RM, Bush JW. Health-related quality of life measurement for evaluation research and policy analysis. *Health Psychol* 1982;**1**:61–80.

Karnofsky DA, Burchenal JH. The clinical evaluation of chemotherapeutic agents against cancer. In: McLeod CM, editor. Evaluation of chemotherapeutic agents. New York: Columbia University Press, 1949.

Loomes G, Mckenzie L. The use of QALYs in health care decision making. *Soc Sci Med* 1989;**28**:299–308.

McKenna S, Hunt SM, 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. *Br J Med Econ* 1993;**6**:13–24.

Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *BMJ* 1990;**300**:230–5.

Nicholl J, Brazier JE, Milner PC, Westlake L, Kohler B, Williams BT, *et al.* Randomised controlled trial of cost-effectiveness of lithotripsy and open cholecystectomy as treatments for gallbladder stones. *Lancet* 1992;**340**:801–7.

O'Brien BJ. Measurement of health-related quality of life in the economic evaluation of medicines. *Drug Inform J* 1994;**28**:45–53.

O'Brien BJ, Buxton MJ, Ferguson BA. Measuring the effectiveness of heart transplant programmes: quality of life data and their relationship to survival analysis. *J Chronic Dis* 1987;**40**(Suppl 1):137S–58S.

Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Q Life Res* 1993;**2**:477–87.

Revicki DA, Wu AW, Murray MI. Change in clinical status, health status, and health utility outcomes in HIV-infected patients. *Med Care* 1995;**33**:AS173–82.

Rutten Van Molken MPMH, Custers F, Van Doorslaer EKA, Jansen CCM, Heurman L, Maesen FPV, *et al.* Comparison of performance of four instruments in evaluating the effects of salmeterol on asthma quality of life. *Euro Resp J* 1995;**8**:888–98.

Spilker B, Molinek FR, Johnston KA, *et al.*, editors. Quality of life bibliography and indexes. *Med Care* 1990;**28**:DS1–DS77.

Stewart AL, Ware J, editors. Measuring functioning and well-being. Durham: Duke University Press, 1992.

Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986;**5**:1–30.

Torrance GW, Thomas WH, Sackett DL. A utility maximisation model for evaluation of health care programs. *Health Services Res* 1972;**7**(2):118–33.

Torrance GW, Zhang Y, Feeny D, Furlong W, Barr R. Multi-attribute preference functions for a comprehensive health status classification system. Hamilton, Ontario: Centre for health economics and policy analysis, McMaster University, 1992.

Tsevat J, Solzan JG, Kuntz KM, Currier JS, Sell RL, Weinstein MC. health values of patients with human-immunodeficiency-virus – relationship to mental health and physical functioning. *Med Care* 1996;**34**:44–57.

Williams A. 'Should QALYs be programme specific?' by Donaldson, Atkinson, Bond, Wright. *J Health Econ* 1989;**8**:485–7 (discussion: 489–91).

Chapter 7

Reviewing the use of preference- and non-preference-based measures of health in economic evaluations published in 1995

Introduction

In earlier chapters of this report we outlined in some detail a number of issues surrounding the use of both preference-based and non-preference-based measures. In this chapter we use the criteria developed in chapter 3 to develop a set of questions which are applied to a number of economic evaluations as identified in a literature search of papers published in 1995. After running the search strategy on various databases a 'screening' process was carried out on abstracts and bibliographic details to identify papers for inclusion in this review. Following the retrieval of potentially includable papers, a number of inclusion criteria were applied to the papers as a final check before a critical review of practice took place. This chapter reviews the criteria previously set out for judging the validity and suitability of the various health measures within the context of economic evaluations of healthcare technology. The various criteria are then synthesised in a way which produces a set of questions which can be applied directly to the economic evaluations considered here. The application of the key questions to the economic evaluations is next described and discussed. Finally in the light of the above analysis some conclusions, recommendations and guidance for future economic evaluations are outlined.

Search strategy

The following terms were used for the search of databases for particular types of economic evaluation published in 1995:

- cost minimisation (picking up both 'minimisation ' and 'minimization')
- cost effective* analys* (picking up both 'analysis' or 'analyses')
- cost utility
- economic evaluation
- quality-adjusted-life-year* (picking up 'year' or 'years')
- qaly* (picking up 'QALY' or 'QALYs')
- cost benefit analy* (picking up both 'analysis' or 'analyses').

Materials were then restricted to English language only, and editorials, letters or news items were subsequently excluded.

The above strategy was taken from MEDLINE, where the designated subject headings 'quality-adjusted-life-years' and 'cost-benefit-analysis' were also used. This strategy was then translated into equivalent strategies for EMBASE, the Science Citation Index, the Social Science Citation Index, Healthstar, CINAHL and NEED. From the health databases all materials were retrieved whilst for the two general science databases materials were manually reviewed for an association with health sciences, for example pharmaceuticals, medical technology, and procedures and techniques. From NEED all materials for 1995 were retrieved, as it was assumed that all the contents of this database were eligible for inclusion in this particular sampling frame.

Inclusion criteria and search results

The search strategy as applied across seven databases produced a total of 1659 papers (see *Table 5*). The identified abstracts and bibliographic details were reviewed according to a set of criteria. To be included in the review each paper had to meet the following criteria:

(a) the study had to be an 'economic' evaluation as defined by Drummond *et al.* (1987) – this implies the following:
  (i) the technology in question had to have a comparator (before and after studies were thus excluded)
  (ii) both costs and consequences (outcomes) of the technology had to be identified and primary data gathered accordingly

TABLE 5 *Number of papers produced by search of the different databases*

Database	No. of papers identified
MEDLINE	538
NEED	123
Other databases – as below Social Science Citation Index CINAHL (nursing) EMBASE Science Citation Index HealthSTAR	998[a]
Total number of articles initially identified in search	1659

[a] *After duplicates from MEDLINE and NEED were taken out*

(b) the outcomes or consequences of technologies had to be assessed using a recognised preference- or non-preference-based measure of HRQoL.

The application of these criteria produced the following results:

- MEDLINE – 21 papers (of 538) meeting the criteria at the initial review stage
- NEED – four papers (of 123) meeting the criteria at the initial review stage
- Other databases – 17 papers (of 998) meeting the criteria at the initial review stage.

A total of 42 papers (approximately 2.5% of the original number) were thus initially identified as economic evaluations in terms of using either preference-based measures (hence CUEs) or non-preference-based HSMs (hence CCAs).

On receipt of the 42 papers the inclusion criteria outlined above were once again strictly applied. As a result of this process the papers were divided into those that were 'in' the final critical review of the use of health measures within economic evaluations and those that were 'out'. The actual numbers in each database meeting the inclusion criteria were as follows:

- MEDLINE – three papers out of 21
- NEED – four papers out of four
- Other databases – six papers out of 17.

Reasons for the resulting high rate of exclusion

At first sight the actual number of evaluations meeting the specified criteria from the papers retrieved may seem low. Some 29/42 (69%) could not be included; the main reason for such a high exclusion rate was an inability to apply the inclusion criteria based on abstracts. Once read, papers which prima facie should be included turned out, for example, not to have actually gathered health data or cost data. These 'economic evaluations' may have relied on literature-based data. Articles without an abstract were obviously much more difficult to judge, and some trade-off between being systematic (ordering all likely papers) and using resources wisely (the cost of gathering what turn out to be non-includable papers) had to be made. Articles which claim in their title to be an economic evaluation (cost-effectiveness studies) often turned out not to be. The NEED database, however, has an advantage for reviewers in that a full review of the paper is given along with 'expert' commentary from a health economist. The result is that all four papers ordered as a result of the search of the NEED database were included in the present review.

The actual reasons for excluding papers, and the number of papers to which they applied, are set out in *Table 6*.

TABLE 6 *Reasons for excluding papers*

Reason for exclusion of paper	No. of papers excluded (x/29)
CEA (b)[a]	1
Hypothetical CUA example (a)	1
Paper used modelling without any primary data collection on costs or health (a)	6
No comparator – not an economic evaluation (a)	5
No cost data collected – not an economic evaluation (a)	3
No HSM in study (b)	5
QALY values used are authors' opinion (b)	2
Article reviews existing evidence only (a)	6

[a] *'(a)' and '(b)' refer to inclusion criteria (see main text)*

Further details of the papers excluded along with the actual reasons for exclusion of each paper are set out in the appendix 5.

The criteria for judging the validity and suitability of HSM instruments from an economist's perspective

In chapter 3 we produced a check-list (see *Box 3*) which can be used to judge the merits of preference-based measures of HRQoL. The three broad criteria which were considered are practicality, reliability and validity. The key idea was that the developers and subsequent users of QALY-type instruments should be able to make some assessment of the merits of that particular instrument. If a QALY-type measure does not pass the test of validity (in particular, empirical validity) then any subsequent decisions based on a CUA are unlikely to lead to improved efficiency in the way resources are deployed.

In chapter 6 we provided a synopsis of the views of health economists on the way in which non-preference-based health status questionnaires have been or could be used in economic evaluations. The relative efficiency of interventions can be compared in situations of clear dominance (where cost and outcomes are superior), where interventions cost the same and outcomes are better on one dimension but no worse on any other, and where outcomes are found to be identical and a cost-minimisation can be performed. The absence of interval properties makes it impossible to undertake CEA when trade-offs are required. In these circumstances non-preference-based HSMs might be useful in a CCA. In this type of analysis a range of costs and consequences of alternative interventions are presented separately with no cost-effect ratio shown.

In seeking to apply the above reasoning to published economic evaluations our experience in reading papers suggested that direct application of the check-list presented in chapter 3 was likely to prove fruitless. For a number of reasons, including limitations of space, authors of economic evaluations do not allude in any great detail, if at all, to the underlying properties of the measure being used. For this reason we have outlined a set of questions to be realistically asked of any published economic evaluation included in this review.

Questions to be applied to published economic evaluations which use HSMs

In light of the above observations we outlined three key questions to be applied to the 13 studies which fitted the inclusion criteria. These questions are presented below, together with the rationale and explanation for each one.

Question 1. Did the paper discuss or address the issue of the validity and/or suitability of a HSM for the particular study question?

When designing an economic evaluation study it should be standard practice to ensure that the HSM chosen will provide reliable estimates of effect and/or be suitable for the particular study setting. This should lead evaluators to ask about the psychometric properties of a disease-specific or generic instrument. In some clinical areas previous work may have suggested that some instrument may be more suitable than others. If this is so, then this should be acknowledged by the authors

When using a QALY instrument, questions regarding the sensitivity of the instrument may be asked. Users of QALY instruments may also seek information regarding the validity of both the health state descriptions used as well as the method of eliciting values (e.g. TTO or SG – see chapter 4 where the various valuation techniques are reviewed).

In both preference- and non-preference-based HSMs this question looks for more than a mere mention of, or reference to, a methodological paper which established an instrument.

Question 2. Were the HSM and the form of economic evaluation chosen compatible?

As stated above, if a study uses a non-preference-based measure then in the majority of cases (scenarios 4–7 in *Table 4*) the only legitimate type of economic evaluation that can be used is a CCA. If a paper was to report a cost per effect ratio (i.e. present a CEA) it would be necessary to show that the measure used has interval measurement properties. In the absence of such evidence this would constitute an incorrect usage of an instrument.

Similarly if a cost–utility framework is adopted, the instrument used to produce QALY gains should be based in utility theory, and address the trade-off between quality and quantity of life. Any attempt to produce QALY values from the NHP for example (an instrument which is not preference-

based and does not factor in the quantity of life gained) would be flawed, and consequently the results should be judged accordingly. QALY values should be estimated rather than 'assumed' or the values of experts adopted ('guessed values').

Question 3. Given the HSM and economic evaluation framework used, were any legitimate conclusions presented in the paper?

In other words, if papers claim to show that one option is more 'cost-effective' than another, can this be substantiated given the health measure used? The validity of any advice, for example on resource allocation offered to decision-making, by authors depends upon the 'quality' of outcomes (and, of course, cost data) gathered. If a study used a sample of five clinicians to arrive at QALY values for treatment results, this may cast some doubt on the usefulness (generalisability) of the findings or any conclusions drawn.

Application of the questions to published studies

As these questions have not previously been set out, it may not be surprising to find that applying them to recently published economic evaluations results in many negative answers. The analysis presented below may be seen in part to be an exploration of the quality of the reporting of studies which claim to be 'economic evaluations', if not of the underlying design features.

The tables presented below provide a synopsis of the results of applying the three questions outlined in the previous section. *Table 7* contains studies classified as CCAs, whilst *Table 8* features studies classified as CUAs. Apart from the application of the three key questions, details of the HSM used are given together with the clinical area being evaluated.

In *Table 7* it can be seen that only four of the nine studies reviewed make any mention of the validity and/or suitability of the HSM instrument used. When this issue is dealt with, then usually only a minimal amount of detail is given. For question 2, six of the studies reviewed can be said unequivocally to have used the HSMs correctly within the CCA framework. This question was not applicable to two studies as they were merely reports of study design and were not at the stage of reporting results. One study (Uyl-de-Groot *et al.*, 1995) used a QALY instrument (EuroQol) but did not make use of this in any cost–utility framework and gave no reason. Another study (Wimo *et al.*, 1995) had mapped Index of Well-Being values from the Global Deteriation Scale but had not presented any cost per QALY calculations. The NHS Centre for Reviews and Dissemination review of this paper had commented that the actual process of mapping Global Deteriation Scale values on to the Index of Well-Being was of dubious merit. The results of applying HSM question 3 are much less clear. It is possible to answer positively without any reservations in only four of the seven papers which reported results. The paper by Johnson *et al.* (1995), which deals with surgery for limb-threatening ischaemia, has two particular problems. Firstly, the paper states that 'we did not know the quality of life scores prior to the onset of limb-threatening ischaemia'. Secondly, the paper claims that limb salvage 'can be the most cost-effective way of managing limb-threatening ischaemia'. This claim looks dubious in that this form of treatment does not appear to dominate the alternative (limb revascularisation) on all the HSM domains recorded. One paper did not state the HSM values found in the study (Knobbe *et al.*, 1995), thus making the paper difficult to use in any form of decision-making.

In the four papers classed as 'cost–utility' studies, two alluded to the validity or suitability of the utility measure used. The QALY instrument used was suited to the framework of CUA in three of the studies. In the study by Mark *et al.* (1995) the method for valuing the health states (the TTO technique) was not clearly explained, and no actual QALY values were presented. This makes any objective judgement of the presentation of results very difficult.

Conclusions: recommendations for future economic evaluations

It is immediately obvious that any analysis of the results of applying the three questions is limited by the small numbers of studies examined. It is somewhat surprising that we found only 13 papers that fitted our criteria. It is true that the inclusion criteria are quite tightly drawn but they are based upon the generally recognised fundamentals of any economic evaluation (e.g. see Drummond *et al.*, 1987). We had some initial scepticism about the low numbers found but having searched seven databases and identified and looked at some 1700 abstracts and bibliographic entries we claim some credibility for our findings.

It is possible, however, in the light of having read the other 29 papers (and indeed drawing from

TABLE 7 *Cost–consequences analysis*

Article authors	Patient group	HSM1	HSM2	HSM3	Key questions		
					Question 1	**Question 2**	**Question 3**
Cottrell *et al.*	Respiratory tract	SIP			No – but original SIP articles referred to	Yes	Yes
Hallstrom *et al.*	Cardiac arrythmia	SF-36	CES-D		Yes	Not applicable	No results yet – interim report
Johnson *et al.*	Limb ischaemia	HADS	Barthel	FAI	Yes – suitability to condition discussed	Yes	? Claims cost-effectiveness ? No baseline HSM results?
Knobbe *et al.*	Mental illness	RLI	SNLAF		Yes	Yes	? HSM values not given (missing table)
Lawrence *et al.*	Hernia surgery	SF-36	EQ-5D (VAS only)		No	Yes	Yes
Prince *et al.*	Tetraplegic care/ rehab	RAND-36	LSI-A	CHART	Yes	Yes	Yes
K Small Aneurysm Trialists	Aortic surgery	MOS SF GHS			No	Not applicable	No results yet – interim report
Uyl-de-Groot *et al.*	Chemotherapy non-Hodgkin's lymphoma	NHP	RSC	KPI	No	Yes – no ratio for EQoL	Yes
Wimo *et al.*	Mental health	GDS	IWB (QALY) mapped from GDS		No	? IWB (QALY) mapped from GDS	No cost per QALY calculated – mapping of GDS to IWB?

CES-D, Centre for Epidemiological Studies Depression Scale; CHART, Craig Handicap Assessment and Reporting Technique; FAI, Frenchay Activities Index; GDS, Global Deterioration Scale; HADS, Hospital Anxiety and Depression Scale; IWB, Index of Well-Being; KPI, Karnofsky Performance Index; LSI-A, Life Satisfaction Index – A; MOS SF GHS, Medical Outcomes Study Short Form General Health Survey; RAND-36, Medical Outcomes Study Short Form General Health Survey; RLI, Resident Lifestyle Inventory; RSC, Rotterdam Symptom Checklist; SNLAF, Social Network Lifestyle Analysis Form

TABLE 8 *Cost–utility analyses*

Article authors	Clinical area	HSM1	HSM2	HSM3	Key questions		
					Question 1	**Question 2**	**Question 3**
Gournay *et al.*	Mental health	QALY Rosser	GHQ	BDI	Yes	Yes	? No present-ation of ratios
Kennedy *et al.*	Lung cancer	QALY TTO			No	Yes	Yes
Kerridge *et al.*	Intensive care	QALY Rosser			Yes	Yes	Yes
Mark *et al.*	Thrombolytic therapy	QALY TTO			No	TTO? Valuation process not clear	? Cost per QALY calculations flawed?

BDI, Beck Depression Inventory; GHQ, General Health Questionnaire

our general experience as readers of economic evaluations) to make some further comments. It appears to be common practice that little or no mention of the validity of HSM instruments used is made. It is highly desirable (and hopefully probable) that consideration of such issues occurs at the design stage of economic evaluations. However, it is also desirable that authors of papers should at least present enough information to allow readers of published papers to gather information on the validity of instruments. More desirable would be some discourse on the HSM chosen. Whilst acknowledging editorial space as a barrier to this, the credibility of economic evaluations and hence their ability to impact on decision-makers would be enhanced if this procedure was followed.

The actual validity and suitability of HSM instruments is likely to be even more obscured in papers which use modelling techniques. Whilst this type of study was excluded from our review, the use of modelling in economic evaluations is increasing (Sheldon, 1996). The debate surrounding modelling is beyond the scope of this report but the question of justifying the choice of HSMs should apply equally in practice. Authors who report the results of modelling work often take parameters from a number of sources and also make assumptions about key values. At the final analysis one is often left with a choice of retrieving all the primary data cited (often not possible due to the way in which the paper is written), or accepting that the author (or the analytical technique used) has enough credibility to be trusted.

Based on our work presented both in this and the preceding chapters, we would recommend that researchers planning economic evaluations pay careful attention to any evidence of the validity of competing HSM instruments, and report these accordingly. Furthermore, that researchers need to be persuaded that using non-QALY measures means that any claims to 'cost-effectiveness' of one technology over another may be severely limited. This guidance is not meant to imply that other aspects of trial and study design are not also important, but, to date, the points made here have not been well recognised. Perhaps the adoption of some of the above in the guidelines used by journals for refereeing economic evaluations would help facilitate change.

References

Cottrell JJ, Openbrier D, Lave JR, Paul C, Garland JL. Home oxygen therapy: a comparison of 2- vs 6-month patient reevaluation. *Chest* 1995;**107**(2):358–61.

Drummond M, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press.

Gournay K, Brooking J. The community psychiatric nurse in primary care: an economic analysis. *J Adv Nurs* 1995;**22**:769–78.

Hallstrom AP, Greene HL, Wyse DG, Zipes D, Epstein AE, Domanski MJ, *et al.* Antiarrhythmics versus implantable defibrillators (avid) – rationale, design, and methods. *Am J Cardiol* 1995;**75**:470–5.

Johnson BF, Evans L, Drury R, Datta D, Morris-Jones W, Beard JD. Surgery for limb threatening ischaemia: a reappraisal of the costs and benefits. *Eur J Vasc Endovasc Surg* 1995;**9**(2):181–8.

Kennedy W, Reinharz D, Tessier G, Contandriopoulos A-P, Traput I, Champagne F. Cost utility of chemotherapy and best supportive care in non-small cell lung cancer. *PharmacoEconomics* 1995;**8**:316–23.

Kerridge RK, Glasziou PP, Hillman KM. The use of 'quality-adjusted life years' (QALYs) to evaluate treatment in intensive care. *Anaesth Intensive Care* 1995;**23**:322–31.

Knobbe CA, Carey SP, Rhodes L, Horner RH. Benefit–cost analysis of community residential versus institutional services for adults with severe mental retardation and challenging behaviours. *Am J Ment Retard* 1995;**99**(5):533–41.

Lawrence K, McWhinnie D, Goodwin A, Doll H, Gordon A, Gray A, *et al.* Randomised controlled trial of laparoscopic versus open repair of inguinal hernia: early results. *BMJ* 1995;**311**:981–5.

Mark DB, Hlatky MA, Califf RM, Naylor CD, Lee KL, Armstrong PW, *et al.* Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *New Engl J Med* 1995;**332**:1418–24.

Prince JM, Manley MS, Whiteneck GG. Self-managed versus agency-provided personal assistance care for individuals with high level tetraplegia. *Arch Phys Med Rehab* 1995;**76**:919–23.

The UK Small Aneurysm Trial Participants. The UK Small Aneurysm Trial: design, methods and progress. *Euro J Vasc Endovasc Surg* 1995;**9**(1):42–8.

Sheldon TA. Problems of using modelling in the economic evaluation of health care. *Health Econ* 1996;**5**:1–13.

Uylde Groot CA, Hagenbeek A, Verdonck LF, Lowenberg B, Rutten FFH. Cost–effectiveness of abmt in comparison with chop chemotherapy in patients with intermediate- and high-grade malignant non-hodgkin's lymphoma (NHL). *Bone Marrow Trans* 1995;**16**:463–70.

Wimo A, Mattson B, Krakau I, Eriksson T, Nelvig A, Karlsson G. Cost–utility analysis of group living in dementia care. *Int J Technol Assess Health Care* 1995;**11**(1):49–65.

Chapter 8

Recommendations

The purpose of this review has been to produce guidance on the use of HSMs in economic evaluation and to suggest avenues for future research. This final chapter includes recommendations on selecting HSMs for use in economic evaluation, on the choice of technique for valuing health and the choice of MAUSs, and on a research agenda.

Guidance on the use of HSMs in economic evaluation

Selection of HSM

We recommend that researchers planning to conduct economic evaluations alongside clinical trials pay careful attention to the check-list in chapter 3 when selecting HSMs.

The purpose of the check-list approach is to provide guidance rather than rules, and to acknowledge disagreements where they exist. The precedent in health economics is the widely used check-list for economic evaluation advocated by Drummond et al. (1987), which was compiled despite disagreements which exist in the literature. The check-list we propose is likely to need up-dating given further theoretical development and the accumulation of more evidence.

Selection of valuation technique

We recommend only choice-based techniques, either SG or TTO, be used to value health states.

We also recommend that SG and TTO values are obtained directly, rather than estimating them from VAS values via a mapping function.

Selection of MAUSs

We recommend either the EQ-5D or the HUI

The HUI-II should be the instrument of choice for children. The leading contenders at the moment for adults are the EQ-5D and the HUI-III. Most UK researchers are likely to favour the EQ-5D on the grounds that it has been more widely used this country and there are UK weights.

Using non-preference-based HSMs in economic evaluation

We recommend that researchers recognise the limitations of non-preference-based HSMs at the design stage of a study.

Preference-based HSMs can only be used to examine the relative efficiency of interventions:

- in situations of clear dominance, that is, where both the cost and outcomes of one intervention are superior
- where interventions cost the same and outcomes are better on one dimension but no worse on any other or
- where outcomes are found to be identical and a cost-minimisation can be performed.

It is not possible to use non-preference-based HSMs to undertake a CEA and hence assess efficiency when trade-offs must be made between dimensions of health and/or cost. Therefore:

We recommend a MAUS be used in all economic evaluations alongside clinical trials.

Research agenda

We limit ourselves to those areas we have reviewed, but, as noted elsewhere in the report (see chapter 2), there are many other important topics in the measurement of healthcare benefits (e.g. the application of HYEs)

Valuation techniques

The following areas of research warrant further work:

- the theoretical foundations of TTO and PTO
- the extent of the violations of the theoretical foundations of each technique
- research to ascertain whether the decisions made as a result of individuals elicited preferences are in line with their actual preferences
- research to develop PTO as a measure of social preferences, and testing its practicality, consistency and empirical validity
- empirical work on mapping from the VAS to TTO and SG at an aggregate level (i.e. for

average health state values), including extensive testing of specification and robustness

Multi-attribute utility scales

The following areas warrant further research:

- Comparing the performance of the EQ-5D and HUI-III. To assist researchers in deciding between the EQ-5D and HUI-III, and indeed whether either is a valid measure of preferences, we recommend that research is undertaken to compare them in terms of the check-list in chapter 3 for assessing descriptive validity (including the construct validity and sensitivity), valuation methods and their empirical validity (against stated and hypothetical preferences).
- Studies to revalue existing MAUSs. Weights from a sample of the UK population need to be estimated using a choice-based valuation technique for (1) the HUI-II, (2) the HUI-III if the HUI-III is found to perform well against the EQ-5D, and (3) the QWB, though whether it this should be a research priority in the UK is doubtful given it is rarely used in the UK.
- The development of new MAUSs. The EQ-5D and HUI may not be suitable for many patient groups, and another measure, perhaps based on a more sensitive classification, is required. This could either be done by (1) developing a new MAUS *de novo*, such as the development of the Australian quality of life measure or (2) estimating preference weights for an existing measure of health status such as the SF-36, though this is only advisable for the most widely used generic and possibly condition-specific measures.
- Research into the valuation of scenarios constructed from HSM data. This approach has the advantage of being able to focus on those aspects of health most relevant to the treatment being evaluated. The methodology for constructing the vignettes from the potentially large volume of descriptive data is not well established and needs to be developed.

Future reviews

The health economics literature on the valuation of healthcare benefits is large and has been growing rapidly since the five literature searches undertaken for this report were done. Much of the research agenda set out above is being addressed, and this will result in parts of this report becoming out of date in the future. Out of the five reviews undertaken for this report, however, only four will need to be considered for renewal. There would seem to be little advantage in further updating the review of the relationship between preference- and non-preference-based measures since the results confirm an earlier review which found only poor correlations between them. At the same time, there are important topics excluded from this report which should be reviewed for the HTA programme in the near future, including the development of HYEs, conjoint analysis and WTP.

The review of the methods of using HSMs in economic evaluation provided the material for two chapters of this report. The first concerned the construction of a check-list for judging measures for use in economic evaluation which exposed a number of areas where there is currently little or no concensus amongst health economists. It would be worth considering renewing this in 5 years time. However, it is not necessary to reexamine the question of how non-preference-based measures should be used in economic evaluation (as addressed in chapter 7) since this has been adequately answered. The review of valuation techniques will need to be updated, but the accumulation of original research, particularly in gaining a better understanding of the meaning behind respondents answers to the tasks and tests of validity, may be slow since this work is of a fundamental nature. It may be 5–8 years before it will be worthwhile updating this review. By contrast there has been a substantial increase in the use of MAUSs in economic evaluations, often with two being used in tandem, and this growing evidence could be used to address a number of outstanding questions about the scales. Therefore the comparison of MAUSs should be updated in the next 3–5 years. Finally, it would be important to repeat the review of studies using health status questionnaires in economic evaluations published in a more recent year to examine whether practice has been improving and to identify scope for further improvement.

Acknowledgements

This study was supported by the NHS R&D Executive's Health Technology Assessment Programme, project number 93/47/08.

We are indebted to the referees for their perseverance in reading the report and the quality of their comments.

Appendix 1

Supplementary data for chapter 3

Papers identified in first search

Anderson JP, Bush JW and Berry CC. Internal Consistency Analysis: a method for studying the accuracy of function assessment for health outcome and quality of life evaluation. *J Clin Epidemiol* 1988;**41**:127–37.

Bakker CH, Rutten van Molken M, van Doorslaer E, Bennett K, van der Linden S. Health related utility measurement in rheumatology: an introduction. *Patient Educ Couns* 1993;**2**0:145–52.

Bakker C, Rutten van Molken M, Hidding A, van Doorslaer E, Bennett K, van der Linden S. Patient utilities in ankylosing spondylitis and the association with other outcome measures. *J Rheumatol* 1994;**21**:1298–304.

Bakker C, Rutten M, van Doorslaer E, Bennett K , van der Linden S. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. *J Rheumatol* 1994;**21**:269–74.

Bakker C, Rutten M, Van SantenHoeufft M, Bolwijn P, van Doorslaer E, Bennett K, *et al.* Patient utilities in fibromyalgia and the association with other outcome measures. *J Rheumatol* 1995;**22**:1536–43.

Bennett K, Torrance G, Tugwell P. Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. *Controlled Clin Trials* 1991;**12**:118S–28S.

Birch S, Gafni A. Ethical dimensions of health technology-assessment – being economic with economic-principles. *Int J Tech Manage* 1995;60–72.

Bleichrodt H. QALYs and HYEs – under what conditions are they equivalent. *J Health Econ* 1995;**14**:17–37.

Bordley RF. Making social trade-offs among lives, disabilities, and cost. *J Risk Uncertainty* 1994;**9**:135–49.

Boyd NF, Sutherland HJ, Heasman KZ, Tritchler DL, Cummings BJ. Whose utilities for decision analysis? *Med Decis Making* 1990;**10**:58–67.

Brazier J, Dixon S. The use of condition specific outcome measures in economic appraisal. *Health Econ* 1995;**4**:255–64.

Brodin H, Persson J. Cost–utility analysis of assistive technologies in the European-Commissions Tide program. *Int J Tech Assess Health Care* 1995;**11**:276–83.

Broome J. QALYs. *J Public Econ* 1993;**50**:149–67.

Brown PB. An alternative to QALYs – saved young life equivalent (save). *BMJ* 1992;**305**:1365.

Buckingham K. A note on HYE (healthy years equivalent). *J Health Econ* 1993;**12**:301–9.

Buckingham K. Economics, health and health economics – HYEs versus QALYs – a response. *J Health Econ* 1995;**14**:397–8.

Carr Hill RA. Assumptions of the QALY procedure. *Soc Sci Med* 1989;**29**:469–77.

Carrhill RA. Background material for the workshop on QALYs – assumptions of the QALY procedure. *Soc Sci Med* 1989;**29**:469–77.

Carr-Hill RA, Morris J. Current practice in obtaining the "Q" in QALYs: a cautionary note. *BMJ* 1991;**303**:699–701.

Carrhill R, Morris J. The Q in QALYs. *BMJ* 1991;**303**:1480.

Coast J. Reprocessing data to form QALYs. *BMJ* 1992;**305**:87–90.

Coast J. Developing the QALY concept – exploring the problems of data-acquisition. *PharmacoEconomics* 1993;**4**:240–6.

Cook J, Richardson J, Street A. A cost-utility analysis of treatment options for gallstone disease – methodological issues and results. *Health Econ* 1994;**3**:157–68.

Culyer AJ. Measuring health: lessons for Ontario. Toronto: University of Toronto Press, 1978.

Culyer AJ, Wagstaff A. QALYs versus healthy year equivalents (HYEs) – a reply to Gafni, Birch and Mehrez. *J Health Econ* 1995;**14**:39–45.

Dolan P, Gudex C, Kind P. Valuing health states: a comparison of methods. *J Health Econ* 1996;**2**:209–32.

Donaldson C, Atkinson A, Bond J, Wright K. Should QALYs be programme-specific? *J Health Econ* 1988;**7**:239–57.

Donaldson C, Atkinson A, Bond J, Wright K. QALYS and long-term care for elderly people in the UK: scales for assessment of quality of life. *Age Ageing* 1988;**17**:379–87.

Donaldson C, Wright K. Program-specific QALYs – a reply. *J Health Econ* 1990;**8**:489–91.

Drewett RF, Minns RJ, Sibly TF. Measuring outcome of total knee replacement using quality of life indices. *Ann R Coll Surg Engl* 1992;**74**:286–9.

Drummond MF. Resource allocation decisions in health care: a role for quality of life assessments? *J Chronic Dis* 1987;**40**:605–19.

Drummond MF. Quality-adjusted life-years. *Lancet* 1987;**i**:1372–3.

Drummond MF. Utility approach to measuring health-related quality of life – discussion. *J Chronic Dis* 1987;**40**:601–3.

Drummond MF, Heyse J, Cook J, McGuire A. Selection of end points in economic evaluations of coronary-heart-disease interventions. *Med Decis Making* 1993;**13**:184–90.

Drummond M, Torrance G, Mason J. Cost-effectiveness league tables: more harm than good? *Soc Sci Med* 1993;**37**:33–40.

Elvik R. The validity of using health state indexes in measuring the consequences of traffic injury for public-health. *Soc Sci Med* 1995;**40**:1385–98.

Eraker SA, Baker S, Miyamoto JM. Parameter estimates for a quality adjusted life year utility model. *Clin Res* 1984;**32**:A294.

Erickson P, Wilson RW. The term years of healthy life – misunderstood, defended, and challenged – a short-term for quality-adjusted life years. *Am J Public Health* 1994;**84**:866.

Essink-Bot ML, Vanroyen L, Krabbe P, Bonsel GJ, Rutten FFH. The impact of migraine on health-status. *Headache* 1995;**35**:200–6.

Evans RW. Quality-adjusted life-years. *Lancet* 1987;**i**:1372.

Feeny DH, Torrance GW. Incorporating utility-based quality-of-life assessment measures in clinical trials. Two examples. *Med Care* 1989;**27**:S190–204.

Ferguson BM, Keown PA. An introduction to utility measurement in health care. *Infect Control Hosp Epidemiol* 1995;**16**:240–7.

Fox Rushby J. Appraising the use of contingent valuation: a note in response. *Health Econ* 1993;**2**:361–2.

French MT, Mauskopf JA. A quality-of-life method for estimating the value of avoided morbidity. *Am J Public Health* 1992;**82**:1553–5.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – I: measurement strategies. *J Clin Epidemiol* 1989;**42**:345–54.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – II: scaling methods. *J Clin Epidemiol* 1989;**42**:459–71.

Fryback DG. QALYs, HYEs, and the loss of innocence. *Med Decis Making* 1993;**13**:271–2.

Gafni A. The quality of QALYs (quality-adjusted-life-years): do QALYs measure what they at least intend to measure? *Health Policy* 1989;**13**:81–3.

Gafni A. Measuring the adverse effects of unnecessary hypertension drug therapy: QALYs vs HYE. *Clin Invest Med* 1991;**14**:266–70.

Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health Serv Res* 1994;**29**:207–24.

Gafni A, Birch S. Searching for a common currency – critical-appraisal of the scientific basis underlying European harmonization of the measurement of health related quality-of-life (EuroQoL(c)). *Health Policy* 1993;**23**:219–28.

Gafni A, Birch S. Guidelines for the adoption of new technologies: a prescription for uncontrolled growth in expenditures and how to avoid the problem. *Can Med Assoc J* 1993;**148**:913–17.

Gafni A, Birch S. Preferences for outcomes in economic evaluation: an economic approach to addressing economic problems. *Soc Sci Med* 1995;**40**:767–76.

Gafni A, Birch S, Mehrez A. Economics, health and health economics – HYEs versus QALYs. *J Health Econ* 1993;**12**:325–39.

Ganiats TG, Miller CJ, Kaplan RM. Comparing the quality-adjusted life-year output of 2 treatment arms in a randomized trial. *Med Care* 1995;**33**:AS245–54.

Gerard K. Cost-utility in practice – a policy makers guide to the state-of-the-art. *Health Policy* 1992;**21**:249–79.

Gerard K, Mooney G. QALY league tables: handle with care. *Health Econ* 1993;**2**:59–64.

Gerard K, Dobson M, Hall J. Framing and labelling effects in health descriptions: quality adjusted life years for treatment of breast cancer. *J Clin Epidemiol* 1993;**46**:77–84.

Glicher SR. Alternative to QALYs – saved young life equivalent (SAVE). *BMJ* 1992;**305**:1365.

Golan EH, Shechter M. Contingent valuation of supplemental health care in Israel. *Med Decis Making* 1993;**13**:302–10.

Goldhirsch A, Gelber RD, Simes RJ, Glasziou P, Coates AS. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *J Clin Oncol* 1989;**7**:36–44.

Grogono AW, Woodgate DJ. Index for measuring health. *Lancet* 1971;Nov 6:1024–26.

Gudex C, Kind P. The Q in QALYs. *BMJ* 1991;**303**:1137.

Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;**118**:622–9.

Hadorn DC. The role of public values in setting health care priorities. *Soc Sci Med* 1991;**32**:773–81.

Hadorn DC, Hays RD, Uebersax J, Hauber T. Improving task comprehension in the measurement of health state preferences. A trial of informational cartoon figures and a paired-comparison task. *J Clin Epidemiol* 1992;**45**:233–43.

Haes JCJM. The integration of quality-of-life and survival – quality-adjusted life years. *Q Life Res* 1993;**2**:60.

Hall J. Best medical practice in practice: measuring efficiency in mammography screening. *Int J Health Plan Manage* 1989;**4**:235–46.

Hall J, Gerard K, Salkeld G, Richardson J. A cost utility analysis of mammography screening in Australia. *Soc Sci Med* 1992;**34**:993–1004.

Harwood R. Reprocessing data to form QALYs. *BMJ* 1992;**305**:424.

Jefferson T, Mugford M, Demicheli V. QALY league tables. *Health Econ* 1994;**3**:205.

Johannesson M. Economic evaluation of hypertension treatment. *Int J Technol Assess Health Care* 1992;**8**:506–23.

Johannesson M. QALYs, HYEs and individual preferences – a graphical illustration. *Soc Sci Med* 1994;**39**:1623–32.

Johannesson M. Quality-adjusted life-years versus healthy-years equivalents – a comment. *J Health Econ* 1995;**14**:9–16.

Johannesson M. QALYs – a comment. *J Public Econ* 1995;**56**:327–8.

Johannesson M. The ranking properties of healthy-years equivalents and quality-adjusted life-years under certainty and uncertainty. *Int J Tech Assess Health Care* 1995;**11**:40–8.

Johannesson M. The relationship between cost-effectiveness analysis and cost–benefit analysis. *Soc Sci Med* 1995;**41**:483–9.

Johannesson M, Jonsson B. Economic evaluation in health care: is there a role for cost–benefit analysis? *Health Policy* 1991;**17**:1–23.

Johannesson M, Pliskin JS, Weinstein MC. A note on QALYs, time tradeoff, and discounting. *Med Decis Making* 1994;**14**:188–93.

Jonsson B. Quality of life – economic aspects. *Scand J Prim Health Care Suppl* 1990;**1**:93–6.

Kaplan RM. Health outcome models for policy analysis. *Health Psychol* 1989;**8**:723–35.

Kaplan RM. Quality of life assessment for cost/utility studies in cancer. *Cancer Treat Rev* 1993;**19** (Suppl A):85–96.

Kaplan RM, Coons SJ. Relative importance of dimensions in the assessment of health-related quality of life for patients with hypertension. *Prog Cardiovasc Nurs* 1992;**7**:29–36.

Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures. *Q Life Res* 1993;**2**:467–75.

Kiebert GM, Stiggelbout AM, Leer JW, Kievit J, de Haes HJ. Test–retest reliabilities of two treatment-preference instruments in measuring utilities. *Med Decis Making* 1993;**13**:133–40.

Kind P, Gudex CM. Measuring health-status in the community – a comparison of methods. *J Epidemiol Comm Health* 1994;**48**:86–91.

Kind P, Dolan P. The effect of past and present illness experience on the valuations of health states. *Med Care* 1995;**33**:AS255–63.

Launois R. La prise en compte des preferences des patients dans les choix de sante individuels et collectifs [Integration of patients' preferences in individual and collective health choices]. *Rev Epidemiol Sante Publique* 1994;**42**:246–62.

Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. *Can Med Assoc J* 1992;**146**:473–81.

Leu RE. Economic evaluation of new drug therapies in terms of improved life quality. *Soc Sci Med* 1985;**21**:1153–61.

Levin LA, Jonsson B. Cost-effectiveness of thrombolysis – a randomized study of intravenous rt-PA in suspected myocardial infarction. *Eur Heart J* 1992;**13**:2–8.

Lipscomb J. Time preference for health in cost-effectiveness analysis. *Med Care* 1989;**27**:S233–53.

Llewellyn Thomas HA, Sutherland HJ, Thiel EC. Do patients' evaluations of a future health state change when they actually enter that state? *Med Care* 1993;**31**:1002–12.

Loomes G. The myth of the HYE. *J Health Econ* 1995;**14**:1–7.

Loomes G, Mckenzie L. The use of QALYs in health care decision making. *Soc Sci Med* 1989;**28**:299–308.

Loomis J, King M. Comparison of mail and telephone-mail contingent valuation surveys. *J Environ Manage* 1994;**41**:309–24.

Mason J, Drummond M, Torrance G. Some guidelines on the use of cost effectiveness league tables. *BMJ* 1993;**306**:570–2.

Mason J, Drummond M. The DH register of cost-effectiveness studies: content and quality. *Health Trends* 1995;**27**:50–6.

Maxwell S. Valuation of rural environmental improvements using contingent valuation methodology: a case study of the Marston Vale Community Forest project. *J Environ Manag* 1994;**41**:385–99.

McGuire A, Hughes D. The economic-evaluation of depression. *Postgrad Med J* 1994;**70**:S14–S22.

McTurk L. A methodological quibble about QALYs. *BMJ* 1991;**302**:1601.

Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and healthy-years equivalents. *Med Decis Making* 1989;**9**:142–9 (erratum: *Med Decis Making* 1990;**10**(2):148–9).

Mehrez A, Gafni A. Evaluating health related quality of life: an indifference curve interpretation for the time trade-off technique. *Soc Sci Med* 1990;**31**:1281–3.

Mehrez A, Gafni A. The healthy-years equivalents: how to measure them using the standard gamble approach. *Med Decis Making* 1991;**11**:140–6.

Miyamoto JM, Eraker SA. Parameter estimates for a QALY utility model. *Med Decis Making* 1985;**5**:191–213.

Miyamoto JM, Eraker SA. A multiplicative model of the utility of survival duration and health quality. *J Exp Psychol Gen* 1988;**117**:3–20.

Morss SE, Lenert LA, Faustman WO. The side effects of antipsychotic drugs and patients' quality of life: patient education and preference assessment with computers and multimedia. *Proc Ann Symp Comput Appl Med Care* 1993;17–21.

Mulley AG Jr. Assessing patients' utilities. Can the ends justify the means? *Med Care* 1989;**27**:S269–81.

Naylor CD, Llewellynthomas HA. Can there be a more patient-centered approach to determining clinically important effect sizes for randomized treatment trials. *J Clinical Epidemiol* 1994;**47**:787–95.

Nease RF Jr, Kneeland T, O'Connor GT, Sumner W, Lumpkins C, Shaw L, *et al.* Variation in patient utilities for outcomes of the management of chronic stable angina. Implications for clinical practice guidelines. Ischemic Heart Disease Patient Outcomes Research Team. *JAMA* 1995;**273**:1185–90.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Planning Manage* 1991;**6**:234–42.

Nord E. Estimating the social value of health care outcomes in terms of saving young life equivalents. *Med Decis Making* 1992;**12**:348P.

Nord E. Methods for quality adjustment of life years. *Soc Sci Med* 1992;**34**:559–69.

Nord E. Toward quality assurance in QALY calculations. *Int J Technol Assess Health Care* 1993;**9**:37–45.

Nord E. The QALY – a measure of social value rather than individual utility? *Health Econ* 1994;**3**:89–93.

Nord E. The person-trade-off approach to valuing health care programs. *Med Decis Making* 1995;**15**:201–8.

Nord E, Richardson J, Macarounas Kirchmann K. Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. *Int J Technol Assess Health Care* 1993;**9**:463–78.

O'Brien BJ. Measurement of health-related quality of life in the economic evaluation of medicines. *Drug Info J* 1994;**28**:45–53.

O'Brien B, Rushby J. Outcome assessment in cardiovascular cost–benefit studies. *Am Heart J* 1990;**119**:740–7.

O'Brien BJ, Buxton MJ, Ferguson BA. Measuring the effectiveness of heart transplant programmes: quality of life data and their relationship to survival analysis. *J Chronic Dis* 1987;**40**(Suppl 1):137S–58S.

O'Connor AM, Pennie RA. Reliability and validity of measures used to elicit health expectations, values, tradeoffs and intentions to be immunized for hepatitis B. *J Clin Epidemiol* 1995;**48**:255–62.

Olsen JA. On what basis should health be discounted? *J Health Econ* 1993;**12**:39–53.

Olsen JA. Persons vs years: two ways of eliciting implicit weights. *Health Econ* 1994;**3**:39–46.

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Med Decis Making* 1994;**14**:9–18.

Rabin R, Rosser RM, Butler C. Impact of diagnosis on utilities assigned to states of illness. *J R Soc Med* 1993;**86**:444–8.

Ranaboldo CJ, Chant ADB. Reprocessing data to form QALYs. *BMJ* 1992;**305**:424.

Rawles J, Light J, Watt M. Loss of quality-adjusted days as a trial end-point – effect of early thrombolytic treatment in suspected myocardial-infarction. *J Epidemiol Commun Health* 1993;**47**:377–81.

Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Q Life Res* 1993;**2**:477–87.

Revicki DA, Simpson KN, Wu AW, LaVallee RL. Evaluating the quality of life associated with rifabutin prophylaxis for Mycobacterium avium complex in persons with AIDS: combining Q-TWiST and multi-attribute utility techniques. *Q Life Res* 1995;**4**:309–18.

Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994;**39**:7–21.

Richardson J, Schwartz S. Quality-adjusted life years: a reply. *Aus J Public Health* 1994;**18**:227–8.

Robinson R. Cost–utility analysis. *BMJ* 1993;**307**:859–62.

Rutten Van Molken MPMH, Custers F, Van Doorslaer EKA, Jansen CCM, Heurman L, Maesen FPV, *et al.* Comparison of performance of four instruments in evaluating the effects of salmeterol on asthma quality of life. *Euro Resp J* 1995;**8**:888–98.

Ruttenvanmolken MPMH, Bakker CH, Vandoorslaer EKA, Vanderlinden S. Methodological issues of patient utility measurement – experience from 2 clinical-trials. *Med Care* 1995;**33**:922–37.

Sayers G. Alternative to QALYs – saved young life equivalent (SAVE). *BMJ* 1992;**305**:1365.

Selai CE, Rosser RM. Good quality quality – some methodological issues. *J R Soc Med* 1993;**86**:440–3.

Siegrist J, Junge A. Background material for the workshop on QALYs – conceptual and methodological problems in research on the quality of life in clinical medicine. *Soc Sci Med* 1989;**29**:463–8.

Smith A. Qualms about QALYs. *Lancet* 1987;**i**:1134–6.

Smith R, Dobson M. Measuring utility values for QALYs: two methodological issues. *Health Econ* 1993;**2**:349–55.

Spiegelhalter DJ, Gore SM, Fitzpatrick R, Fletcher AE, Jones DR, Cox DR. Quality-of-life measures in health-care. 3. Resource-allocation. *BMJ* 1992;**305**:1205–9.

Sutherland HJ. Assessing patients' preferences. *Med Decis Making* 1995;**15**:286–7.

Till JE, Sutherland HJ, Meslin EM. Is there a role for preference assessments in research on quality of life in oncology? *Q Life Res* 1992;**1**:31–40.

Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986;**5**:1–30.

Torrance GW. Utility approach to measuring health-related quality of life. *J Chronic Dis* 1987;**40**:593–600.

Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care* 1989;**5**:559–75.

Torrance GW, O'Brien B. An interview on utility measurement. *J Rheumatol* 1995;**22**:1200–2.

Tsevat J, Goldman L, Soukup JR, Lamas GA, Connors KF, Chapin CC, Lee TH. Stability of time-tradeoff utilities in survivors of myocardial infarction. *Med Decis Making* 1993;**13**:161–5.

Turner RP, Lustig SP. Using quality-of-life measures to produce QALYs in clinical-trial – a comparison of 3 methods. *Q Life Res* 1994;**3**:55–6.

Verhoef CG, Maas A, Stalpers LJA, Verbeek ALM, Wobbes T, Van Daal WAJ. The feasibility of additive conjoint measurement in measuring utilities in breast cancer patients. *Health Policy* 1991;**17**:39–50.

Wade DT. The Q in QALYs. *BMJ* 1991;**303**:1136–7.

Weinstein MC. A QALY is a QALY – or is it? *J Health Econ* 1988;**7**:289–90.

Wilkinson G, Williams B, Krekorian H, Mclees S, Falloon I. QALYs in mental-health – a case-study. *Psych Med* 1992;**22**:725–31.

Williams A. Economics of coronary artery bypass grafting. *BMJ* 1985;**291**:326–9.

Williams A. Quality-adjusted life-years. *Lancet* 1987;**i**:1372.

Williams A. 'Should QALYs be programme specific?' by Donaldson, Atkinson, Bond and Wright. *J Health Econ* 1989;**8**:485–7 (discussion: 489–91).

Williams C, Coyle D, Gray A, Hutton J, Jefferson T, Karlsson G, *et al.* European-school-of-oncology advisory report to the commission of the European communities for the Europe against cancer program cost- effectiveness in cancer care. *Eur J Cancer* 1995;**31A**:1410–24.

Appendix 2

Supplementary data for chapter 4

Empirical review/listing

We have provided a listing of studies, identified in the review of health state valuation techniques, to provide outline detail on the performance of the techniques and to offer some guidance on the application of techniques to particular samples, for example patient/disease groups. The listing is not a fully systematic one, that is, not all studies identified presenting empirical data are contained in the table presented below. We have included those studies which have applied the techniques in a clinical setting and those methods-based studies which report empirical findings which are able to inform on the practical application of the techniques (even where a convenience sample has been used). Where studies have been found to report previously presented data, with only a slight change in focus, they have not been listed in the table (overleaf). Outline information relating to 120 references is presented.

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•					Cairns *et al.* (1996)	Antenatal screening for cystic fibrosis (fetal loss versus information) (n = 52 women from general population)	Methodological article – valuing benefits from screening	SG. Administered via interview	Health states were given high values. Values of '*p*' used in SG were often small	Individuals may have difficulty in dealing with these small numbers
•					Clarke and Badaway (1991)	Appendicitis (n = 66)	Hypothetical operation for suspected appendicitis	SG (with props). Administered via interview	No comment made on performance	Variation in elicited utilities were wider than expected (relative to perforation and self-limiting abdominal pain with no operation)
•					Hatziandreu *et al.* (1994)	Recurrent depression	Modelling effects of maintenance therapy with sertraline versus dothiepin (utility values for four depression-related health states from two physician panels)	SG	No comment made on performance	
•					Kavanagh *et al.* (1996)	Chronic heart failure (n = 30)	Non-randomised controlled trial of long-term benefits of aerobic training in patients with chronic heart failure	SG	SG showed a 14% increase from initial assessment of quality of life	
•					Nichol *et al.* (1996)	Angina (n = 41)	Methodological focus – relationship between cardiac capacity and patient symptom-specific utilities	SG	No comment made	
•					Rabin *et al.* (1993)	Selection of diagnostic conditions including disability, destress and discomfort/pain. Convenience sample of health volunteers (n = 42)	Study to assess the impact of diagnostic information on values people assign to selected health states. Methodological paper	SG. Administered via interview. Health state duration of 1 year	SG found to be straight-forward to administer and few subjects reported difficulty. 100% completion	Found consistently low scores attached to mental conditions
•					Ramsey *et al.* (1995)	Lung transplantation (waiting list patients, n = 21; post-transplant patients, n = 23)	Cross-sectional study	SG. Administered via interview	Completion rate 95%; 2 of 46 unable to complete due to difficulties understanding. Patients were willing and able to respond to SG questions despite great physical and mental burdens of disease. SG can be used in multicentre evaluations	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•					Revicki (1992)	Chronic renal disease		SG	12 of 73 respondents were unable to complete SG task or gave inconsistent responses	
•					Shackley and Cairns (1996)	Antenatal screening (female general population, n = 52 woman)	Methodological focus – evaluating benefits of a screening programme	SG	No comment made on SG	
•					Thompson (1986)	Rheumatoid arthritis	Assessment of rheumatoid arthritis health state using willingness to undertake risky intervention involving a hypothetical cure (or WTP for cure) (n = 247). Experimental study	SG (and WTP). Health state was a hypothetical cure	98% estimated a maximum acceptable risk using the SG method. Rates of response were high for all educational levels and slightly correlated with education	Author undertook regression analysis view, showing that maximum acceptable risk grew with disease duration and a decline of maximum acceptable risk with co-morbidities
•					Tousignant *et al.* (1994)	Obstructive sleep apnoea (n = 19)	Nasal continuous air pressure. Cross-sectional study	SG	Test–retest after 2–4 weeks, intraclass correlation of 0.78 indicated a high level of reliability. Tested against symptoms, SG was meaningfully related to disease-specific measure	
	•				Ashby *et al.* (1994)	Breast cancer; patients, nurse, hospital doctors, general practitioners, university staff (n = 138)	Treatment for breast cancer (methodological study – do different groups of respondents give different valuations?)	TTO (with props) Administered via interview. Chronic health states	TTO method was practical and acceptable to subjects. Good test–retest results at 3–6 weeks. Mean TTO values consistent with respondent rank ordering	Patients gave higher values for health states; age and sex influenced valuation with TTO
	•				Churchill *et al.* (1990)	Anaemia patients receiving dialysis (n = 118)	Recombinant human erythropoeitin versus placebo. Randomised controlled trial	TTO	TTO valuations found no significant difference in health states (the SIP did find improvements in quality of life)	
	•				Churchill *et al.* (1991)	End-stage renal failure (n = 47; TTO, n = 7)	Methodological – comparison of indices of quality of life in haemodialysis patients	TTO	TTO was not responsive; no correlation with clinical measures	
	•				Dolan *et al.* (1996a)	General population study (n = 3395)	EQ-5D valuation survey	TTO. Administered via interview	TTO valuations were highly consistent. TTO shown to be feasible and acceptable (only 1.3% of responses excluded from analysis). Reliability tested on 221 respondents – mean intraclass correlation: 0.73 (median 0.79)	Valuation for severe states affected by age and sex of respondent. Conclude that it is feasible to use TTO to elicit valuation from the general public

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
	•				Fryback *et al.* (1993)	Population-based study of eye disease prevalence and risk factors (n = 1356)	Study of health status and HRQoL	TTO (with props; life expectancy duration). Administered via interview	Good evidence of acceptability, 38 did not complete the TTO task. However over 50% of respondents would trade no life years for remediation of their health problems	TTO results expressed as a percentage of remaining life expectancy
	•				Glasziou *et al.* (1994)	Acute myocardial infarction (n = 714)	Thrombolytic agents post-myocardial infarction – cohort study	TTO (self-completed)	TTO completion rate 91%, no other observations	Large number of respondents were not prepared to give up any time in the TTO task, although 66% indicated they were already in full health
	•				Handler *et al.* (1997)	Convenience sample (n = 86)	Methods study	TTO. Automated interview	100% completion of TTO task	Authors considered TTO in connection with 'locus of control', and find some effect on TTO response
	•				Irvine *et al.* (1995)	Ulcerative colotis (n = 43) and Crohn's disease (n = 51)	Quality of life measurement study	TTO	47% of patients refused to trade-off any life expectancy despite experiencing a large number of problems	Authors find that TTO is useful for detecting large-scale changes over time or large between group differences
	•				Johnson *et al.* (1996)	Cytomegalovirus in HIV-positive patients (n = 80)	Cross-sectional study patient preferences for intravenous ganciclovir treatment	TTO. Administered via interview	Completion rate for TTO 100%	
	•				Krahn *et al.* (1994)	Prostate cancer (sample: n = 10 medics)	Screening for prostate cancer. Decision-analytic study	TTO ('gambler' automated method)	TTO scores (utilities) were similar to those found in other prostate cancer work	
	•				Kreibich *et al.* (1996)	Knee replacement (n = 67)	Total knee replacement	TTO (with props). Administered via interview	TTO completion rate high.	TTO recorded pre- and post-treatment effects lower than health status measures – TTO was less able to discriminate change than the SF-36. But did record significant change in health status
	•				Krumins *et al.* (1988)	Benign prostatic hyperplasia (n = 20)	Transurethral resection of the prostate surgery decision-analytic model	TTO	Feasible to use TTO in this patient group. TTO values found to be moderately correlated with symptom severity	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
	•				Laupacis *et al.* (1996)	Renal transplantation (n = 136)	Renal transplantation versus dialysis cohort study	TTO	Significant difference in TTO score pre- and post-transplant found (18 and 24 months later)	
	•				Laupacis *et al.* (1993)	Osteoarthritis (n = 185)	Total hip replacement (with or without cement) randomised controlled trial	TTO	TTO found 2 year improvement in patients' own health. Responses to hypothetical scenarios were consistent in terms of pre- and post-operative valuations	
	•				Laupacis *et al.* (1991)	Haemodialysis (n = 118)	Effect of erythropoietin on quality of life and exercise capacity, randomised controlled trial	TTO		
	•				McLeod *et al.* (1995)	(Unspecified) benign or malignant neoplasm (n = 25)	Whipple procedure cross-sectional study	TTO	No performance detail reported	
	•				McLeod *et al.* (1991)	Ulcerative colitis (n = 113)	Surgery for ulcerative colitis	TTO	Authors state that TTO 'seems to be valid in measuring quality of life this group of patients'	
	•				Molzhan *et al.* (1996)	End-stage renal disease (n = 215)	Methodological – model explaining quality of life in end-stage renal disease patients	TTO. Duration of 30 years. Administered via interview	Test–retest (6 weeks) $r = 0.85$	
	•				Oldridge *et al.* (1993)	Acute myocardial infarction and depression (n = 201)	Cardiac rehabilitation	TTO	No performance detail reported	
	•				Provenzale *et al.* (1997)	Chronic pancolitis (n = 22)	Protocolectomy and ileal pouch	TTO	TTO correlated with the SIP physical subscale ($r = -0.55$) and with SF-36 physical subscales ($r = 0.43$–0.51) ($p < 0.05$). TTO could discriminate important health status changes (physical functioning) in this group of patients	
	•				Perez *et al.* (1997)	Advanced cancer (n = 93)	Methodological – quality of life measures	TTO (self-administered)	TTO (not strongly related to quality of life in this study) provides a measure of patients attitudes to their health state. 90% of respondents reported good understanding of TTO questions. But only 76% stated that TTO was easy to use, compared with 88–93% in the Spitzer quality of life measure	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
	•				Russelll *et al.* (1992)	End-stage renal disease (n = 27)	Renal transplantation	TTO	Test–retest (6–8 weeks) 'found to be highly reproducible' – figures not given	
	•				Smith *et al.* (1993)	Colonic carcinoma – Dukes stage C (n = 16)	Adjuvant chemotherapy	TTO	No performance detail presented	
		•			Adang *et al.* (1996)	Insulin-dependent diabetes mellitus patients undergoing pancreas–kidney transplantation. Combined pancreas–kidney transplantation (n = 17) versus combined pancreas–kidney transplantation with loss of pancreas (pancreas failed) (n = 5)	Multicentre prospective longitudinal design	VAS (10 cm line from worst possible quality of life to best possible quality of life; rating own health states)	No problems with completion or acceptability to patients reported	No detail offered on valuation task
		•			Wilkinson *et al.* (1997)	Subclinical thiamine deficiency in an elderly population (65 years or over) (n = 96)	Randomised double-blind treatment trial – oral thiamine or placebo	VAS (10 cm – quality of life assessment)	No detail give regarding VAS operation or performance	Predominantly clinical paper. No detail given of the VAS methods. VAS quality of life measures were able to show evidence of subjective treatment effects with improvement in quality of life
		•			Bethoux *et al.* (1996)	Spouses of stroke patients (6 months after stroke) (n = 9)	Observational study, assessing quality of life in stroke patients and spouses	VAS (10 cm – quality of life assessment)	Shows a statistically significant correlation with Barthel index scores ($r = 0.78$, $p = 0.026$)	No general discussion of methods
		•			Carlson *et al.* (1995)	Home parenteral nutrition (n = 37)	Observational study, assessing quality of life in patient group	VAS to assess quality of life from 0 (the lowest) to 10 (the highest)	No details reported by authors. Study reports that there was a statistically significant correlation between the objective quality of life measure employed and VAS scores ($r = 0.6$, $p = 0.001$)	No general discussion of methods
		•			Cook *et al.* (1996)	Percutaneous transluminal angioplasty for unilateral claudication (n = 29)	Observational study, examing quality of life before and after percutaneous transluminal angioplast	VAS (EuroQoL questionnaire)	VAS identified a significant pre- and post-treatment increase in quality of life	No details reported
		•			Curry *et al.* (1996)	Left ventricular hypertrophy with essential mild-to-moderate hypertension (n = 112)	Prospective multicentre open study, testing the effect of indapimide (a diuretic) on the patient group	VAS (no details reported)	VAS scores indicated a statistically significant improvement in quality of life (overall well-being)	No further detail reported

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
		•			Brooks *et al.* (1991) (EuroQoL Group)	Representative sample of the Swedish population (16–84 years)	Health state valuation, pilot study. Core and restricted core sets of health states	VAS (thermometer type; EuroQoL six dimensions). Postal questionnaire	Some reluctance with response and some incomplete response problems	Early version of the VAS from EuroQoL Group
		•			Chen *et al.* (1996)	Chronic stable angina, with no history of revascularisation (n = 55)	Cross-sectional interview study of health status and valuation. Health states were (1) current health, i.e. stable chronic angina, and (2) angina-free state	VAS (verbal rating scale method). Interviews to rate health between 0 (death) and 100 (no health problems and feeling really good)	Anticipated gain varied widely. Patients appeared to believe that even relief of angina would still leave their health far from perfect	Study results suggest that anginal severity (clinical measures) may not be that closely correlated with health state preferences
		•			Essink-Bot *et al.* (1990)	Random sample of the Dutch general population (n = 200; responders, n = 112)	Health state valuation study	VAS (EuroQoL-style postal questionnaire, six-dimension early EuroQoL descriptor). 14 health states valued (duration 1 year)	88 non-responders. Of the 112 responding 32 reported problems completing the questionnaire, five returned blank questionnaires and 21 had not understood the task (mean time 20.3 min). 7 respondents valued 'dead' equal to or higher than 50. Evidence of good inter-rater reliability (two states presented twice, and individual correlations were high)	Authors find that 'rating of health states on a VAS by postal questionnaire appears to be feasible. However, inappropriate response of those returning questionnaire did occur (20%) and was related to age and level of education'
		•			Fernandez *et al.* (1989)	Chemotherapy for non-small cell lung cancer (n = 31)	Clinical trial involving patients undergoing two courses of chemotherapy	VAS (100 mm horizontal line with anchor points). Overall quality of life	Quality of life was reported improved by 75% of patients. No correlation was found between Karnofsky performance scores and VAS quality of life scores (r = 0.10, p = 0.58)	Anchor points were described in a symptom-related manner
		•			Fries and Ramey (1997)	Rheumatoid arthritis	Quality of life assessment study (n = 663; follow-up sample, n = 43)	VAS (EuroQoL thermometer variant). Baseline, follow-up sample at 6 months	Not much detail given about VAS methods employed, e.g. application. Correlations between the global assessment scale, as used in Health Assessment Questionnaire, and VAS feeling thermometer technique for estimating quality of life were high. Change scores are similarly very highly correlated. Feeling thermometer and global analogue scale, r = –0.68, n = 663, $p \leq 0.001$; retest, n = 43, r = –0.72, $p \leq 0.0001$	This study seems to suggest that the VAS scale is closely linked to HSM scores. Authors mention that the VAS was more likely to pick up co-morbidities than a specific rheumatoid arthritis scale

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
		•				Rheumatoid arthritis	Pilot study of patient preferences regarding non-steroidal anti-inflammatory drugs related adverse gastrointestinal events and their prophylaxis (n = 30)	VAS ((CR) with anchors at 0 (immediate death) and 100 (full health for life)). Interviews; valuing six health states in differing hypothetical positions	Respondents' comments indicated the interview to be acceptable and feasible. Mean time for completion 56 min. Use mean ratings for health states	Authors do not report much detail on valuation methods or performance
		•			Glaspy (1997)	Wide variety of non-myeloid malignancies being treated with chemotherapy	Observational, open-label study (n = 1498 patients; completers = 987). Impact of epoetin alpha therapy for anaemia on quality of life	VAS (linear analogue scale, horizontal, 100 mm line). Overall quality of life, anchored at 'as low as could be' and 'as high as it could be'	Not much detail reported surrounding valuation methods	Identified statistically significant improvement in quality of life before and after treatment, the effect size was supported by other status measures undertaken. The magnitude of the increase in quality of life was related to the magnitude of the haemoglobin increase for all response categories
		•			Grunberg *et al.* (1996)	Antiemetic control (supportive care) in patients undergoing chemotherapy	Pilot study to evaluate quality of life improvements associated with antiemetic control, using patients familiar with chemotherapy (n = 30)	VAS (horizontal 100 mm scale ('terrible to wonderful'). Administered by an oncology nurse; reviewing quality of life over a period 3–4 weeks	Strange results between valuation method and quality of life measures	Authors comment on valuation and see rating scale as simplest, with immediate values
		•			Glaspy *et al.* (1997)	Patients with malignancies undergoing cytotoxic chemotherapy (n = 2342; 1498 gave quality of life scores)	Open-label study to assess treatment using epoetin alpha	VAS (linear analogue scales for overall quality of life; horizontal scale, 100 mm). Quality of life anchors were 'worst possible' and 'best possible'	1498 patients had baseline and completion quality of life data	Not much commentary on health state valuation. Results reported within a clinical paper

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
		•			Barr *et al.* (1995)	Paediatric oncology, e.g. acute lymphoblastic leukaemia, Wilms's tumour, neuroblastoma	Elicitation of health state values relating to the patient group. Not undertaken alongside a clinical trial	VAS (feeling thermometer)	No details reported, reader referred to details reported elsewhere	Also uses SG method but authors do not report on any comparison of the two methods used
		•			Gudex *et al.* (1996)	General population (n = 3395)	EuroQoL valuation study	VAS (15 health states). Duration 10 years. Administered via interview	Test–retest (n = 221, 10 weeks after initial survey) intraclass correlation 0.78. VAS said to be 'acceptable to a wide range of respondents'. Only 3.2% of responses were excluded from analysis. Rate of inconsistency reported at 2.5%	Study found social class and education had a significant effect on valuations
		•			Hopman-Rock *et al.* (1997)	Community-living elderly with pain in the hip or knee (osteoarthritis) (n = 306)	Testing relationship between increased pain and lower quality of life scores	VAS	The VAS showed lower quality of life in people with more chronic pain ($p < 0.001$). Suggestive that the VAS is sensitive to change in this patient group	
		•			Kwa *et al.* (1996)	Ischaemic stroke hospitalised patients (n = 129)	Investigation of the role of cognitive impairment in stroke patients	VAS	75% of sample (97/129) managed to complete the VAS. The VAS failed to detect impact of cognitive impairment on patients quality of life. Authors state that a 'minimal use of cognitive function and communication skills are necessary for VAS'	
		•			Sculpher *et al.* (1996)	Menorrhagia (n = 196; response rate 155/196 = 79%)	Randomised controlled trial surgical treatment for menorrhagia	VAS	The VAS showed no significant difference in health states produced by treatments. No other comments	
		•			Silvertssen *et al.* (1994)	Open heart surgery patients (n = 1127)	Clinical study containing health state valuation	VAS. Postal questionnaire	VAS scores: –91% of patients had a higher score post-operatively (mean of 30 months later) ($p < 0.001$). The VAS adequately used by 95% of respondents	
		•			Sutherland *et al.* (1983)	Outpatients receiving radiotherapy for a variety of malignant diseases	Methodological study	VAS (linear analogue scale)	Values for health states measured by VAS do not always correspond to the strictly rational expectations of utility theory	Evidence of framing/context effects

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
		•			Ure *et al.* (1994)	Patients with oesophageal atresia who underwent colon interposition (n = 8)	Retrospective follow-up study (22 years)	VAS. Administered via interview	VAS result reported but no comment given on performance	Used 100 mm scale with end-points of 'lowest quality' and 'highest quality'
			•		Murray and Lopez (1997)	Measures of disability in a general population	Study to quantify disability for inclusion in health policy debates. Used nine groups of individuals, consisting of participants from 25 countries	PTO. Administered in a series of group exercises (across regions). Group exercises contained 8–12 people and lasted 10 hours	Provided disability severity weights. Only outline information reported	
			•		Nord (1993b)	Convenience sample (n = 10)	Pilot study – methodological	PTO. Administered via interview	Respondents made 14 pairwise comparisons. Respondents understood the task, but found it difficult to choose precise equivalence numbers. Strong correlation between estimated and observed equivalence numbers	Equivalence numbers for program evaluation. Advocates the SAVE as a unit of measure
			•		Nord *et al.* (1993)	Random sample of Norwegian population (n = 102) and a convenience sample of students and nurses (n = 384; randomly selected from a university and a hospital)	Methods study assessing validity of four scaling instruments using PTO	PTO. Self-administered postal questionnaire (used EQ-5D descriptors)	Response rate: 28.2% in Norway and 27% in Australia	Study assess social valuation only. Considers health state scaling instruments
				•	Rosser and Kind (1978)	20 patients from either medical or psychiatric wards plus 20 nurses, 10 doctors and 20 healthy volunteers (n = 70)	Valuation study; the 29 health states in the Rosser and Kind matrix	ME	Test–retest reliability (as measured by percentage agreement) 97.2%; inter-rater reliability 88%	
				•	Haig *et al.* (1986)	Patients admitted for general surgery (n = 159)	Valuation study; valuing health states with three dimensions – dysfunction, discomfort and prognosis (based on work carried out previously by Patrick *et al.* (1973))	ME. Administered via interview. Absence of dysfunction and discomfort = 0	Measurement error – intrarater reliability coefficients of 0.96–0.98 produced	High correlation between results and results of earlier study by Patrick *et al.* (1973)
•	•				Clarke *et al.* (1997)	Gaucher's disease (low blood cell count) (n = 109)	Observational study – comparison of methods for presenting health states Treatment with alglucerase	SG,TTO (multi-media – computer-generated visuals)	Retest (done with n = 52) after 1–3 weeks after initial valuation work showed SG had lowest mean absolute difference. SG also had lowest intraclass correlation coefficient	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•	•				de-Wit *et al.* (1995) (conference abstract only)	Lung cancer patients 3–24 months post-treatment (n = 15)	Lung cancer treatments (design not specified)	SG, TTO. Administered via interview	SG and TTO can be used (i.e. utility measurement is feasible) in patients with lung cancer. SG and TTO detected differences in perceived quality of life between patients differing in health status as measured by the Karnofsky scale	
•	•				Dolan *et al.* (1996b)	Population study (n = 335)	Comparison of SG and TTO – props and no props versions of both	SG, TTO	Found good levels of completion with both methods – SG with props and without props at 5.3 and 4.4% of states unvalued, respectively, and TTO with props and without props at 0.8 and 4.2% of states unvalued, respectively	
•	•				Gage *et al.* (1996)	Stroke and stroke prophylaxis	Quality of life study	SG, TTO. Interviews using computer-based method	Good test–retest result for stroke utilities 0.67–0.92. 13 of 83 interview excluded due to difficulties with the questions (does not specify if cases were specific to an instrument) – 84% completion. Evidence of consistency – utilities reflected expected ordinal rankings of stroke utilities (mild/moderate/major). No significant difference between SG and TTO utilities	Stroke utilities – wide variation. Warfarin/aspirin utilities – near to one. Large variations in utilities – 'sickness like beauty is in the eye of the beholder'
•	•				Reed *et al.* (1993)	Hypercholesterolaemia (n = 35)	Cholesterol-lowering therapy	TTO, SG	Test–retest results 'comparable to other studies'. Individual responses were stable: TTO, $r = 0.74$; SG, $r = 0.82$. Good correlation found between individual TTO and SG in first and second interviews	
•	•				Stiggelbout *et al.* (1994)	Testicular cancer disease-free patients (n = 30)	Methodological study – valuing four health states relevant to testicular cancer	SG, TTO	Some problems with completion due to hypothetical nature of the questions. SG scores higher than TTO scores. Certainty equivalent-adjusted TTO scores are equal to SG scores	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•		•			Bakker *et al.* (1994b)	Chronic rheumatic diseases – ankylosing spondylitis (n = 72) and fibromyalgia (n = 86)	Ankylosing spondylitis: randomised controlled trial considering non-steroidal anti-inflammatory drugs. Fibromyalgia: study of the effect of fitness training and biofeedback	VAS (thermometer) and SG (with props). Did not use death – used a two-stage process involving a severe marker state. Values for own health: mild health state, moderate state (RS), severe state (SG)	SG produced higher values than the VAS. Both methods proved feasible. Reliability good in both methods – intraclass correlation coefficients: SG, r = 0.77–0.79; VAS, r = 0.70–0.95	States described using six dimensions, using adaptaion of McMaster Utility Measurement Questionnaire. Conclude that utility measurement is sensitive to method used. Study does not compare in detail – does not discuss validity, sensitivity, etc., in any detail
•		•			Bakker *et al.* (1994a)	Ankylosing spondylitis (n = 59)	Randomised controlled trial of supervised group physical therapy versus exercise at home	RS, SG (two-step with props). Patients valued four health states: three marker states plus own health	Both RS and SG methods of health state valuation appeared feasible and reliable. 1 week and 9 month retest results were good. RS values correlated better with health status measures used and disease activity measures used than did SG values	
•		•			Bakker *et al.* (1995)	Fibromyalgia (n = 73)	Randomised controlled trial of therapeutic effect of low-impact fitness training and biofeedback training	VAS, SG (using props; Maastricht Utility Measurement). Administered via interview. Test–retest showed poor reliability. RS did not correlate well with SG (r = 0.14)		
•		•			Bass *et al.* (1994)	Interviewed 40 general medical patients without gallstones regarding treatment for symptomatic gallstones	Study to estimate patient preferences for gallstone-related treatments and outcomes. Main intervention – extracorporeal shock wave lithotripsy	VAS, SG (with props). Administered via interview. VAS (horizontal scale: 0 = immediate death and 100 = perfect health)	Mean preference values in the SG group were highly correlated with the mean preference values in the RS group (Spearman coefficient = 0.90, p = 0.0001). RS values lower than SG values. Both methods performed well on reliability (intrarater) when repeat values were analysed. But a consistent and substantial difference between values derived by an RS and those derived by SG	Even though both scaling techniques yielded reliable results, the two methods clearly did not produce equivalent scale values. SG values were highly correlated with those in the rating scale group

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•		•			Boyd *et al.* (1990)	Rectal cancer (n = 68)	Decision-analytic study selection of surgery (colostomy) or radiation therapy	SG,VAS (10 cm linear analogue scale). Administered via interview	Test–retest examined (2 weeks to 12 months follow-up). Reproducibility was satisfactory (all correlations significant at 5%)	
•		•			Chouinard and Allbright (1997)	Schizophrenia (n = 135)	Randomised controlled trial – multicentre trial. Treatment with risperidone (health states rated by 100 psychiatric nurses)	SG,VAS (linear analogue scale)	Health state utilities for mild, moderate and severe schizophrenia reported. No comments made on performance of instruments	
•		•			Coley *et al.* (1996)	Community-acquired pneumonia (n = 159)	US Pneumonia Patient Outcome Research Team	VAS, SG (with props). Administered via interview	SG values higher than VAS ($p < 0.001$)	
•		•			Goossens *et al.* (1996)	Fibromyalgia (n = 131)	Randomised controlled trial educational/cognition interventions in fibromyalgia	VAS, SG (Maastricht Utility Measure)	RS values 25–35% lower than SG values. Authors suggest that 10% steps for this patient group may be too large: they would not accept a 10% chance of dying when offered treatment. SG as used here not suitable for chronic patients	SG values thus subject to a ceiling effect and has little chance of capturing improvements in this condition
•		•			Goodwin *et al.* (1988)	Small cell lung cancer (patients receiving treatment similar to trial, n = 7; health professionals familiar with treatments, n = 14)	Randomised controlled trial of treatments for small cell lung cancer (retrospective economic evaluation)	VAS, SG	The instruments appeared to be able to differentiate between the six health states valued	
•		•			Hayman *et al.* (1997)	Early stage breast cancer (n = 97)	Conservative surgery and radiation therapy	VAS, SG (with props)	Substantial inter-respondent variability found in utility scores	
•		•			Hutton *et al.* (1996)	Metastatic breast cancer (values from 30 oncology nurses acting as proxy for patients)	Taxied (drug therapy) decision-analytic model	VAS, SG (with props)	Health state ranking was consistent with *a priori* expectations. High degree of consistency in results between countries	
•		•			Kupperman *et al.* (1997)	Pregnant women (n = 121)	Prenatal diagnosis – chorionic villus sampling versus amniocentesis (methods study health state valuation)	VAS, SG (with props)	Standard deviations were higher for VAS scores than for those obtained from SG	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•		•			Llewellyn-Thomas *et al.* (1982)	Cancer patients (n = 64)	Radiotherapy (health state valuation methodology)	SG, VAS	Methods found to be reliable and acceptable. Evidence of a good level of consistency, 54 of 64 respondents gave a rank ordering in line with expected rank order. Good performance on completion. SG values greater than VAS values	Scores for health states seemed to be influenced by the procedures used to elicit values, by health state description techniques, and by sequence of rating methods used. Evidence of violations of EUT – internal inconsistency
•		•			Lenert *et al.* (1997)	Schizophrenia (n = 22 plus 41 healthy volunteers)	Side-effects of antipsychotic drugs	VAS, SG (computerised survey)	All respondents successfully completed tasks. Some evidence that SG values were inconsistent with rank order previously given by respondents. Preferences can be elicited from patients with severe mental illness	
•		•			Morss *et al.* (1994)	Schizophrenia (n = 33) plus psychiatrists (n = 5)	Side-effects of antipsychotic drugs (clozapine)	VAS, SG (multi-media technique – see also Lenert *et al.*, 1997)	SG and VAS instruments were found to be feasible and acceptable. SG and VAS scores showed 76% consistency. VAS values lower than SG values for same health states	
•		•			O'Brien and Viramontes (1994)	Chronic lung disease (n = 102)	Methodological – main focus on WTP	VAS, SG (with props)	SG scores higher than VAS scores; both methods showed a clear and significant gradient between disease severity groups. These instruments are capable of being used in this disease area. Test–retest reliability (4 weeks) intraclass correlation 0.61 for RS and 0.82 for SG	
•		•			O'Brien *et al.* (1990)	Chronic rheumatic disease (patients with ankylosing spondylitis, n = 100)	Methodological focus – patients willingness to accept risk of mortality in drug treatment for rheumatic disease with a hypothetical new drug	SG, VAS	SG gave similar results to those found in a previous study of RA thus some indication of reliability of SG. Convergent validity between SG and the NHP shown	
•		•			Rutten-van Molken (1995)	Moderate asthma (n = 107)	Effects of salmeterol or salbutamol on quality of life in asthma patients (testing sensitivity and construct validity of measures)	RS, SG (with props)	SG failed to detect change in quality of life following drug treatment as indicated by disease-specific instrument and RS. RS performed better than SG in terms of construct validity	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•		•			Revicki *et al.* (1995)	HIV (infected patients, n = 160)	Methodological focus – comparison of psychometric health measures and health utility scales (longitudinal study)	SG, RS	SG did not discriminate between three groups of patients (poor responsiveness). RS did detect changes in clinical symptoms	
•		•			Revicki *et al.* (1996)	Schizophrenia (n = 49) plus care-givers	Valuing five different outcomes as experienced by schizophrenics (CR) and physician rated states using SG and CR	SG,VAS	SG could not be used with schizophrenics; they could use the VAS (CR). SG utilities correlated with patient preferences. SG score obtained from psychiatrists were higher than their CR scores	
•		•			Rutten-van Molken (1995b)	Fibromyalgia (n = 85) and ankylosing spondylitis (n = 144)	Randomised controlled trials of treatment for conditions (main focus methodological issues of patient utility measurement)	VAS, SG (with props)	Good level of consistency with methods; 21% of respondents gave at least one inconsistent response; however, about 70% of these fell within the standard error of measurement.	SG scores consistently higher than VAS (CR) scores. Reliability measured by 3 month retest: intraclass correlation 0.24–0.37 for VAS, and 043–0.70 for SG. Utilities from chained SG were significantly higher than those from basic reference gambles
•		•			Revicki (1992)	Chronic renal disease with anaemia (n = 73)	Randomised controlled trial of r-HuEPO versus placebo	SG (with props), VAS	SG scores consistently greater than VAS (CR) scores for end-stage renal disease with dialysis and severe anaemia needing transfusions. It was found that 19% of sample could not complete the SG valuation exercise compared with 1.3% for the VAS	
•		•			Torrance *et al.* (1996)	Valuation of HUI-II		VAS, SG (with props)	Four health states valued by SG and VAS; SG values greater than VAS in all cases	
	•	•			Daly *et al.* (1993)	Women likely to experience or who were experiencing menopausal symptoms (n = 63)	Assessment of quality of life effects caused by menopausal symptoms	TTO – 5 year time frame. VAS (RS) (numerical scale, 0–10), end-points death and normal health. Administered via interview	Did/could not test reproducibility within the study. Responses found to be internally consistent. Using kappa scores, the two methods produced results that were poorly related but not contradictory	Some evidence of response problems. Some interesting comments to illustrate that people view things differently, e.g. one person sees a score of 7 as 'quite good', another sees 7 as 'not too good'. Severe symptoms seen by some as the same as dead, by others as 'an irritant'

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
	•	•			Detsky *et al.* (1986)	Long-term home parenteral nutrition (n = 37)	Modelling survival of long-term home parenteral nutrition	VAS, TTO	Only 1/37 failed to complete TTO, TTO was used successfully in measuring quality of life in long-term home parenteral nutrition patients. VAS scores were not significantly different from TTO scores	
	•	•			Ferraz *et al.* (1993)	Rheumatoid arthritis (n = 25) plus rheumatologists (n = 25)	Corticosteroid therapy (prednisone) – utility approach to evaluate risks and benefits of therapy	TTO, VAS. Administered via interview	All respondents completed tasks successfully. TTO scores higher than VAS for all three scenarios valued. Correlation of TTO and VAS: 0.675 in physician group and 0.518 in patients	
	•	•			Gabriel *et al.* (1994)	Rheumatoid arthritis (n = 57)	Low-dose misoprostol therapy – decision-analytic model (cost–utility)	TTO, VAS	Reliability (4 weeks after initial interview using small group): Pearson coefficient 0.63 for TTO, and 0.89 for VAS. TTO scores higher than VAS scores	
	•	•			Llewellyn-Thomas *et al.* (1993)	Laryngeal cancer (n = 66)	Radiation therapy. Methodological – do health state valuations remain constant when patients alter experience those health states?	TTO, VAS	Found that TTO and VAS scores did remain stable when those health states were experienced by patients; TTO values exceeded VAS	
	•	•			Robinson *et al.* (1997)	Population sample, subsample of MVH study (n = 43)	Methodological paper examining differing VAS and TTO responses; quantitative and qualitative data	VAS, TTO. Administered via interview	Respondents indicated that TTO values were a better indication of true preferences	When using the VAS, respondents tended to ignore the duration of the health state. TTO states had a 'threshold tolerability' below which states have to fall for respondents to be willing to trade any time at all. Thus, TTO may not be suitable for use in certain clinical settings. Younger respondents (in TTO) find the worse than dead scenario less plausible than older

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
	•	•			Swan *et al.* (1997)	Peripheral vascular disease (n = 30)	X-ray and magnetic resonance angiography	VAS, TTO. Telephone interview	Considerable variability was present in TTO responses. No comment on the VAS	Modified TTO for shorter morbidity. Authors state that with adaptation TTO could be suitable for assessing procedural morbidity in imaging procedures
	•	•			Tsevat *et al.* (1996)	HIV (n = 139)	Methodological – relationship between health values and health status (0–6 months)	TTO, VAS	TTO correlated moderately with other quality of life measures. TTO scores exceeded VAS	
	•	•			Tsevat *et al.* (1993)	Myocardial infarction (n = 67)	Change in health status of myocardial infarction survivors over 18 months	TTO, VAS. Duration at 5 years. Administered via interview	TTO did not correlate with change in functioning of patients. Change in TTO non-significant; suggestion that TTO is not suitable for this patient group	41% of respondents unwilling to trade any time
	•	•			Nord (1991)	General population	EuroQoL tariff valuation. Comparison of values for health states from respondents in Norway compared with three other countries	VAS. A mixture of administration methods	VAS and PTO (equivalent number) techniques yielded the same rankings (with the exception of one pairing). Directly elicited equivalence numbers (PTO) were consistently higher than the equivalence numbers implied by the VAS values	Valuing eight health states (instead of 16 as in other countries) lowered difficulty of task. Qualitative evidence presented to suggest that the VAS was not being used as an equal interval scale
		•		•	Kaplan *et al.* (1979)	Student volunteers (n = 65)	Methods, valuation study – comparing the relationship between VAS (CR) and the unbounded form of ME. Study extends work of Patrick *et al.* (1973)	ME, VAS	ME judged to be inappropriate as a measurement for a health status index. ME scores compressed to lower end of scale near death	
		•		•	Sintonen (1981)	Convenience sample of healthy volunteers (n = 43) and surgical patients (n = 77)	Methods paper, examining the use of ME	ME, VAS	The majority of respondents found the questions acceptable (no specific data). No statistically significant difference between the two methods	This study uses the techniques to value individual dimensions of health and proposes an additive model of valuation. Preliminary methods study
•	•	•			Bleichrodt *et al.* (1997a)	Convenience samples from Stockholm (n = 80) and Erasmus, Rotterdam (n = 92)	Experimental study to examine the performance of SG, TTO and VAS QALYs	SG, TTO, VAS. Eight health states. Multitask interview across groups of ten	Show that in the situation of no discounting, the correlation between predicted ranking and direct ranking was significantly higher for TTO than for VAS and SG QALYs. No significant differences were observed between VAS and SG QALYs. All methods performed well	Authors state It may be that for normative/prescriptive reasons, which are more relevant in health economics and medical decision-making, one wishes to use SG QALYs

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•		•			Bleichrodt *et al.* (1997a)	Convenience samples from Stockholm (n = 80) and Erasmus, Rotterdam (n = 92)	Same empirical study as above. This paper reports on tests of theoretical foundation of the VAS	VAS, SG. Interview	Report experimental test, to examine the interpretation of the VAS as being based on a 'measurable value function' – reject measurable value function argument; find VAS values depend on the number of health states that are preferred and less preferred	Conduct experiment to assess the existence of a stable relationship between VAS and SG valuation, i.e. functional form suggested by Torrance – no evidence to support stable relationship
•	•	•			Bosch and Hunink (1996)	Intermittent claudication (peripheral arterial disease)	Multicentre randomised controlled trial. Comparing relationship between descriptive health status measures and valuational measures	SG, TTO, VAS. Postal questionnaire followed by telephone interview	Test–retest reliability good. 11% refused to answer TTO and SG questions, SG yielded highest utility scores	
•	•	•			Busschbach *et al.* (1994)	Cystic fibrosis (n = 6; 3 post-transplant patients and 3 waiting list patients)	Comparative empirical health state valuation pilot study. Quality of life measurement before and after bilateral lung transplantation	VAS, SG, TTO (chronic health states). Administered via interview	SG produced higher values for health states. Not clear which produced the 'real utilities'. Concluded that it is feasible to measure quality of life with SG, TTO and the VAS. Patients were more willing to give up life expectancy in TTO, than take risk in SG	
•	•	•			Giesler *et al.* (1996) (abstract only)	Advanced prostate cancer (n = 52)	Treatment for advanced prostate cancer (not a clinical study)	VAS, SG, TTO	RS outperformed SG and TTO in ability to differentiate health states. All methods produced inconsistency in valuing health states when compared with initial ranking of health states by respondents	
•	•	•			Hornberger *et al.* (1992)	Renal disease, patients on haemodialysis, (n = 58)	Comparative empirical study of methods to assess well-being	SG, TTO, VAS. Administered via interview	The VAS correlated more closely with health status than did SG or TTO. Large variability demonstrated. VAS scores greater than SG, but not significantly	
•	•	•			Nease *et al.* (1996)	Global health (general population sample, n = 83)	Assessment of global health	VAS, SG, TTO (automated utility assessment)	Inconsistent answers produced by 10% of sample. Automated utility assessment instrument is feasible for assessing utility of overall health	

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•	•	•			Nease *et al.* (1995)	Angina pectoris (n = 220)	Case series – health state values for symptoms of angina to be used in management decisions for treating angina	VAS, TTO, SG (computerised interview 'U' titre method)	Reliability tested (2 week retest), found utilities to be reasonably reliable and valid measures of patient preferences for angina symptoms. Concluded that utility assessment is possible in chronic stable angina.	
•	•	•			Patrick *et al.* (1994)	Convenience sample, nursing home residents (n = 15) and well adults (n = 38)	Methods study, valuing states worse than death	VAS (CR), TTO, SG (with props)	100% completion amongst well adults. 78% completion amongst nursing home residents. Respondents did not find SG and TTO to be more burdensome than the VAS (CR). Found well adults can cope with the cognitive burden of valuing states worse than death	
•	•	•			Torrance (1976)	Various health states assessed in a general population sample, four groups (total, n = 380)	Study to compare methods of health state valuation – methodological	SG, TTO, VAS. Administered via interview	All techniques acceptable. Subjects found TTO easier than SG, and SG easier than the VAS (CR). TTO and SG expensive and time-consuming whilst VAS less costly and quicker. Test–retest reliability at 1 year: SG, r = 0.53; TTO, r = 0.62; VAS, r = 0.49	The VAS gives values that are significantly different from the other two techniques. Torrance discusses transformation of scores using a power function
•	•	•			Van der Donk *et al.* (1995)	Laryngeal cancer (patients, clinicians and public, n = 39)	Surgery or radiation therapy	SG, TTO, VAS (with props)	Experienced some problems with acceptability of tasks. For SG, 11 of 39 had missing values; 40% rated TTO as the most difficult of the methods; 21% considered SG to be the easiest of the methods used. VAS (RS) scores lower than SG or TTO	
•	•	•			Wolfson *et al.* (1982)	Secondary care for stroke patients. Physicians, therapists, patients, family members (n = 60)	Assessed cost-effectiveness of various home and institutional care programs available to stroke patients	SG, TTO, VAS. Administered via interview	SG values higher than TTO, which were higher than the VAS. The VAS and TTO had greater correlation than did either technique with SG. Found SG to be most difficult of the techniques to administer and found VAS and TTO comparable in terms of ease of administration	Suggests effects of gambling aversion. Comments that if a choice had to be made between the techniques used, TTO would seem the most promising

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

continued

Technique[a]					Study	Patient group	Study type/intervention	Instrument	Performance	Comments
S	T	V	P	M						
•	•	•			Zug *et al.* (1995)	Skin disease (psoriasis) (n = 87)	Methotrexate therapy	RS, TTO, SG (computer-based 'U' titre)	RS scores did not correlate with SG and TTO values. SG and TTO not significantly different	
•		•	•		Prades (1997)	Convenience sample (students, n = 30)	Methodological paper – testing PTO over SG and the VAS on ability to make resources allocation decisions. Prediction of allocation decisions directly elicited from the public	SG, PTO, VAS. Administered via interview. Used four generic health states (EQ-5D)	Considered social preferences. VAS scores did not reflect preferences for established priorities. Compared performance of methods on the basis of predictive power against expected ordering, and found the VAS to be poor. SG and two forms of PTO were similar whilst a third method of PTO administration outperformed SG	Develops methodology of PTO, using three variants, and develops methods of social preference assessment
	•	•		•	Gudex *et al.* (1993)	327 members of public interviewed at home	Study revaluing the Rosser matrix and comparing three scaling methods	ME, VAS, TTO. Administered via interview	High degree of consensus between ME, VAS (CR), and TTO in the ranking of states. Consistency of methods reported; ME had three reversals of logical orderings (of 28), TTO had six reversals of logical ordering and VAS (CR) had seven reversals, after controlling for factorial design	ME scale most affected by interviewer bias
		•	•	•	Patrick *et al.* (1973)	Students (n = 231) plus health-leaders (New York State Health Commission) (n = 232)	Methodological paper. Generating health state values for the Index of Well-Being (later known as the QWB)	ME, VAS (CR), PTO (equivalence). Top of scale bounded as well day = 1000	Convergent validity established between ME and CR, instruments shown to be reliable (intrarater coefficients of 0.74–0.83; inter-rater reliability was given as 0.75–0.79). Comment that PTO (equivalence) proved complex	
•	•	•	•		Ubel *et al.* (1996)	Convenience sample (economics students, n = 53). Considered both chronic and fatal health states	Methodological study – tests whether methods for eliciting utilities capture public values for healthcare rationing	VAS, SG, TTO: valued three chronic states. PTO: assessing six rationing choices. All administered via a written survey	TTO and SG scores were higher than VAS scores. VAS, SG and TTO all produced the same ordering of states. 11 of 53 respondents were excluded from the PTO task due to inconsistent responses. In 49 of a total of 252 PTO rationing choices, subjects thought it would take an infinite number of people treated in the less severe condition to equal the benefit of treating ten in the severe condition	Respondents implied PTO numbers differed from direct PTO results. Concluded that utility values are not easily translated into social policy, i.e. rationing decisions

[a] *S, standard gamble; T, time trade-off; V, visual analogue scale; P, person trade-off; M, magnitude estimation*

Discounted (10% rate per annum) and undiscounted TTO health state value

Years in full health (*x*) equivalent to 10 years (*t*) in *hi*	Undiscounted TTO value (i.e. (column 1)/*t*)	Discounted years	Cumulative discounted years	Discounted TTO value (i.e. (column 4)/ (10 discounted years))	Difference
1	0.1	1	1	0.15	0.05
2	0.2	0.91	1.90	0.28	0.08
3	0.3	0.83	2.74	0.41	0.11
4	0.4	0.75	3.49	0.52	0.12
5	0.5	0.68	4.17	0.62	0.12
6	0.6	0.62	4.79	0.71	0.11
7	0.7	0.56	5.35	0.79	0.09
8	0.8	0.51	5.87	0.87	0.07
9	0.9	0.47	6.3	0.94	0.04
10	1	0.42	6.76	1	0

Health state valuation, review references

* Denotes reference identified by bibliographic/ *ad hoc* searches.

Anonymous. Double blind dose-response study of zidovudine in AIDS and advanced HIV infection. Nordic Medical Research Councils' HIV Therapy Group. *BMJ* 1992;**304**:13–17.

Anonymous. Acute pain management in adults: operative procedures... quick reference guide for clinicians. *Dermatol Nurs* 1994;**6**:257–65.

Anonymous. New QOL measure needed for adults with growth hormone deficiency. *Drugs Ther Perspect* 1995;**5**:14–16.

Adang EM, Engel GL, van Hooff JP, Kootstra G. Comparison before and after transplantation of pancreas–kidney and pancreas–kidney with loss of pancreas – a prospective controlled quality of life study. *Transplantation* 1996;**62**:754–8.

Ashby J, O'Hanlon M, Buxton MJ. The time trade-off technique: how do the valuations of breast cancer patients compare to those of other groups? *Q Life Res* 1994;**3**:257–65.

Bakker C, van der Linden S. Health related utility measurement: an introduction. *J Rheumatol* 1995;**22**:1197–9.

Bakker CH, Rutten van Molken M, van Doorslaer E, Bennett K, van der Linden S. Health related utility measurement in rheumatology: an introduction. *Patient Educ Couns* 1993;**20**:145–52.

Bakker C, Rutten M, van Doorslaer E, Bennett K, van der Linden S. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. *J Rheumatol* 1994;**21**:269–74.

Bakker C, Rutten M, van Santen Hoeufft M, Bolwijn P, van Doorslaer E, Bennett K, *et al.* Patient utilities in fibromyalgia and the association with other outcome measures. *J Rheumatol* 1995;**22**:1536–43.

Barer D. Assessment in rehabilitation. *Rev Clin Gerontol* 1993;**3**:169–86.

Barr RD, Feeny D, Furlong W, Weitzman S, Torrance GW. A preference-based approach to health-related quality-of-life for children with cancer. *Int J Ped Hematol/Oncol* 1995;**2**:305–15.

Bass EB, Steinberg EP, Pitt HA, Griffiths RI, Lillemoe KD, Saba GP, *et al.* Comparison of the rating scale and the standard gamble in measuring patient preferences for outcomes of gallstone disease. *Med Decis Making* 1994;**14**:307–14.

Ben Zion U, Gafni A. Evaluation of public investment in health care. Is the risk irrelevant? *J Health Econ* 1983;**2**:161–5.

Bethoux F, Calmels P, Gautheron V, Minaire P. Quality of life of the spouses of stroke patients: a preliminary study. *Int J Rehabil Res* 1996;**19**:291–9.

Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. *Med Decis Making* 1997;**17**:208–16.

Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. *J Health Econ* 1997;**16**:155–75.

Bosch JL, Hunink MG. The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication. *Med Decis Making* 1996;**16**:217–25.

Bowe TR. Measuring patient preferences: rating scale versus standard gamble. *Med Decis Making* 1995;**15**:283–5.

Boyd NF, Sutherland HJ, Heasman KZ, Tritchler DL, Cummings BJ. Whose utilities for decision analysis? *Med Decis Making* 1990;**10**:58–67.

Brazier JE, Williams BT, Nicholl JP, Milner P, Westlake L, Ross B, *et al.* Randomised controlled trial of open cholecystectomy and lithotripsy for gallstones: cost effectiveness one year post intervention abstract. *Abstr Int Soc Technol Assess Health Care* 1992;21–2.

Brazier JE, Harper R, Thomas K, Usherwood T. Deriving a preference-based single index measure from the SF-36. *Annu Meeting Int Soc Technol Assess Health Care* 1996;**12**:50.

Brooks RG, Jendteg S, Lindgren B, Persson U, Bjork S. EuroQoL: health-related quality of life measurement. Results of the Swedish questionnaire exercise. *Health Policy* 1991;**18**:37–48.

Buckingham JK, Birdsall J, Douglas JG. Comparing three versions of the time tradeoff: time for a change? *Med Decis Making* 1996;**16**:335–47.

Buckingham K. A note on HYE (healthy years equivalent). *J Health Econ* 1993;**12**:301–9.

Buckingham K, Drummond N. A theoretical and empirical classification of health valuation techniques. HESG (Summer), Strathclyde, 1993.*

Bult JR, Hunink MG, Tsevat J, Weinstein MC. Heterogeneity in the relationship between the time trade-off and Short Form-36 for HIV-infected and primary care patients. *Annu Meeting Int Soc Technol Assess Health Care* 1995;**11** (abstract 8).

Bult JR, Bosch JL, Hunink MG. Heterogeneity in the relationship between the standard-gamble utility measure and health-status dimensions. *Med Decis Making* 1996;**16**:226–33.

Cairns J, Shackley P, Hundley V. Decision making with respect to diagnostic testing: a method of valuing the benefits of antenatal screening. *Med Decis Making* 1996;**16**:161–8.

Chen AY, Daley J, Thibault GE. Angina patients' ratings of current health and health without angina: associations with severity of angina and comorbidity. *Med Decis Making* 1996;**16**:169–77.

Cher DJ, Miyamoto J, Lenert LA. Incorporating risk attitude into Markov-process decision models: importance for individual decision making. *Med Decis Making* 1997;**17**:340–50.

Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizu A, *et al.* Measurement of quality of life in end-stage renal disease: the time trade-off approach. *Clin Invest Med* 1987;**10**:14–20.

Churchill D, Keown P, Laupacis A, Muirhead N, Sim D, Slaughter D, *et al.* Association between recombinant human erythropoietin and quality of life and exercise capacity of patients receiving haemodialysis. *BMJ* 1990;**300**:573–8.

Churchill DN, Wallace JE, Ludwin D, Beecroft ML, Taylor DW. A comparison of evaluative indices of quality of life and cognitive function in hemodialysis patients. *Control Clin Trials* 1991;**12**:159S–167S.

Clarke AE, Goldstein MK, Michelson D, Garber AM, Lenert LA. The effect of assessment method and respondent population on utilities elicited for Gaucher disease. *Q Life Res* 1997;**6**:169–84.

Clarke JR. A scientific approach to surgical reasoning. V. Patients' attitudes. *Theor Surg* 1991;**6**:166–76.

Clarke JR, Badawy SB. Acute pain over the appendix. A model of the surgical decision. *Ann Chir* 1991;**45**:279–83.

Cohen BJ. Assigning values to intermediate health states for cost-utility analysis: theory and practice. *Med Decis Making* 1996;**16**:376–85.

Cohen J. Preferences, needs and QALYs. *J Med Ethics* 1996;**22**:267–72.

Coley CM, Li YH, Medsger AR, Marrie TJ, Fine MJ, Kapoor WN, *et al.* Preferences for home vs hospital care among low-risk patients with community-acquired pneumonia. *Arch Intern Med* 1996;**156**:1565–71.

Cook TA, O'Regan M, Galland RB. Quality of life following percutaneous transluminal angioplasty for claudication. *Eur J Vasc Endovasc Surg* 1996;**11**:191–4.

Culyer AJ, Wagstaff A. QALYs versus HYEs. *J Health Econ* 1993;**12**:311–23.

Curry CL, Robinson H, Brown R, Olivan J, Sami M, Honos G, *et al.* Regression of left ventricular hypertrophy in patients with essential hypertension. Results of 6 month treatment with indapamide. *Am J Hypertens* 1996;**9**:828–32.

Daly E, Gray A, Barlow D, McPherson K, Roche M, Vessey M. Measuring the impact of menopausal symptoms on quality of life. *BMJ* 1993;**307**:836–40.

de Charro FT, de Wit GA, Merkos M. Utility measurements in an end-stage renal disease (ESRD) population. *Annu Meeting Int Soc Technol Assess Health Care* 1996;**12**:39.

de Charro FT, van Busschbach J, van-Hout BA. Some considerations about negative values for quality of life indices. *Annu Meet Int Soc Technol Assess Health Care* 1996;**12**:9.

de Wit GA, Busschbach JJ, de Charro FT, Aaronson NK, van-Zandwijk N. Utility measurement in patients with lung cancer. *Annu Meeting Int Soc Technol Assess Health Care* 1995;**11** (abstract 66).

Detsky AS, McLaughlin JR, Abrams HB, L'Abbe KA, Whitwell J, Bombardier C, *et al.* Quality of life of patients on long-term total parenteral nutrition at home. *J Gen Intern Med* 1986;**1**:26–33.

Dolan P. The effect of experience of illness on health state valuations. *J Clin Epidemiol* 1996;**49**:551–64.

Dolan P. Modelling valuations for health states: the effect of duration. *Health Policy* 1996;**38**:189–203.

Dolan P, Gudex C. Time preference, duration and health state valuations. *Health Econ* 1995;**4**:289–99.

Dolan P, Kind P. Inconsistency and health state valuations. *Soc Sci Med* 1996;**42**:609–15.

Dolan P, Sutton M. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. *Soc Sci Med* 1997;**44**:1519–30.

Dolan P, Jones Lee M, Loomes G. Risk – risk versus standard gamble procedures for measuring health state utilities. *Appl Econ* 1995;**27**:1103–11.

Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. *Health Econ* 1996;**5**:141–54.

Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996;**15**:209–31.

Dominitz JA, Phillips-Bute B, Provenzale D. Utility assessment of individual preferences for colorectal cancer screening. *AHSR FHSR Annu Meeting Abstr Book* 1995;**12**:55–6.

Donnelly S, Walsh D. Quality of life assessment in advanced cancer. *Palliat Med* 1996;**10**:275–83.

Elvik R. The validity of using health state indexes in measuring the consequences of traffic injury for public health. *Soc Sci Med* 1995;**40**:1385–98.

Essink Bot ML, Bonsel GJ, van der Maas PJ. Valuation of health states by the general public: feasibility of a standardized measurement procedure. *Soc Sci Med* 1990;**31**:1201–6.

Ferguson BM, Keown PA. An introduction to utility measurement in health care. *Infect Control Hosp Epidemiol* 1995;**16**:240–7.

Fernandez C, Rosell R, Abad Esteve A, Monras P, Moreno I, Serichol M, *et al.* Quality of life during chemotherapy in non-small cell lung cancer patients. *Acta Oncol* 1989;**28**:29–33.

Ferraz MB, Quaresma MR, Goldsmith CH, Bennett K, Atra E. Corticosteroids in patients with rheumatoid arthritis: utility measurements for evaluating risks and benefits. *Rev Rheum Engl Ed* 1994;**61**:240–4.

Fischhoff B. Value elicitation: is there anything in there? In: Hechter M, Nadel L, editors. The origin of values, pp.187–214. New York: Alice De Gruyer, 1993.*

Friedman L. Commentary on measuring the impact of menopausal symptoms on quality of life. *ONS Nurs Scan Oncol* 1994;**3**:7 (original article: Daly E *et al. BMJ* 1993;**307**:836–40).

Fries JF, Ramey DR. "Arthritis specific" global health analog scales assess "generic" health related quality-of-life in patients with rheumatoid arthritis. *J Rheumatol* 1997;**24**:1697–702.

Froberg DG, Kane RL. Methodology for measuring health-state preferences – II: scaling methods. *J Clin Epidemiol* 1989;**42**:459–71.

Fryback DG, Lawrence WF Jr. Dollars may not buy as many QALYs as we think: a problem with defining quality-of-life adjustments. *Med Decis Making* 1997;**17**:276–84.

Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K, *et al.* The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making* 1993;**13**:89–102.

Gabriel SE, Campion ME, O'Fallon WM. Patient preferences for nonsteroidal antiinflammatory drug related gastrointestinal complications and their prophylaxis. *J Rheumatol* 1993;**20**:358–61.

Gabriel SE, Campion ME, O'Fallon WM. A cost–utility analysis of misoprostol prophylaxis for rheumatoid arthritis patients receiving nonsteroidal antiinflammatory drugs. *Arthritis Rheum* 1994;**37**:333–41.

Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health Serv Res* 1994;**29**:207–24.

Gage BF, Cardinalli AB, Owens DK. The effect of stroke and stroke prophylaxis with aspirin or warfarin on quality of life. *Arch Intern Med* 1996;**156**:1829–36.

Giesler RB, Brody BA, Ashton CM, Geraci JM, Souchek J, Wray NP. Evaluating the validity of patient preference measures in the absence of a gold standard. *AHSR FHSR Annu Meeting Abstr Book* 1996;**13**:163–4.

Gill TM. Quality of life assessment: values and pitfalls. *J R Soc Med* 1995;**88**:680–2.

Gillis JC, Goa KL. Streptokinase: a pharmacoeconomic appraisal of its use in the management of acute myocardial infarction. *PharmacoEconomics* 1996;**10**:281–310.

Glaspy J. The impact of epoetin alfa on quality of life during cancer chemotherapy: a fresh look at an old problem. *Semin Hematol* 1997;**34**:20–6.

Glaspy J, Bukowski R, Steinberg D, Taylor C, Tchekmedyian S, VadhanRaj S. Impact of therapy with epoetin alfa on clinical outcomes in patients with nonmyeloid malignancies during cancer chemotherapy in community oncology practice. *J Clin Oncol* 1997;**15**:1218–34.

Glasziou PP, Bromwich S, Simes RJ. Quality of life six months after myocardial infarction treated with thrombolytic therapy. AUS-TASK Group. Australian arm of International tPA/SK Mortality Trial. *Med J Aust* 1994;**161**:532–6.

Goldstein MK, Clarke AE, Michelson D, Garber AM, Bergen MR, Lenert LA. Developing and testing a multimedia presentation of a health-state description. *Med Decis Making* 1994;**14**:336–44.

Gonzales JJ, McNeil M, Schulman K, Epstein S, Goldstein D, Goldberg R. Assessing patient preferences for depression health states. *AHSR FHSR Annu Meeting Abstr Book* 1994;**11**:12–13.

Goodwin PJ, Feld R, Evans WK, Pater J. Cost-effectiveness of cancer chemotherapy: an economic evaluation of a randomised trial in small cell lung cancer. *J Clin Oncol* 1988;**6**:1537–47.

Goossens ME, Ploem HV, Rutten MP, Winter J, Leidl RM. A cost-effectiveness analyses of behavioral rehabilitation therapies for chronic low back pain; preliminary results. *Abstr Int Soc Technol Assess Health Care* 1993;**9**:102.

Goossens ME, Rutten van Molken MP, Leidl RM, Bos SG, Vlaeyen JW, Teeken Gruben NJ. Cognitive-educational treatment of fibromyalgia: a randomized clinical trial. II. Economic evaluation. *J Rheumatol* 1996;**23**:1246–54.

Gore JM, Granger CB, Simoons ML, Sloan MA, Weaver WD, White HD, *et al.* Stroke after thrombolysis. Mortality and functional outcomes in the GUSTO-I trial. Global use of strategies to open occluded coronary arteries. *Circulation* 1995;**92**:2811–18.

Gough IR, Furnival CM, Schilder L, Grove W. Assessment of the quality of life of patients with advanced cancer. *Eur J Cancer Clin Oncol* 1983;**19**:1161–5.

Grunberg SM, Boutin N, Ireland A, Miner S, Silveira J, Ashikaga T. Impact of nausea/vomiting on quality of life as a visual analogue scale-derived utility score. *Support Care Cancer* 1996;**4**:435–9.

Grunberg SM, Groshen S, Steingass S, Zaretsky S, Meyerowitz B. Comparison of conditional quality of life terminology and visual analogue scale measurements. *Q Life Res* 1996;**5**:65–72.

Gudex C, Kind P, van Dalen H, Durand MA, Morris J, Williams A. Comparing scaling methods for health state valuations – Rosser revisited. Discussion paper 107. York: Centre for Health Economics, University of York, 1993.*

Gudex C, Dolan P, Kind P, Williams A. Health state valuations from the general public using the visual analogue scale. *Q Life Res* 1996;**5**:521–31.

Hadorn DC. Setting health care priorities in Oregon: cost-effectiveness meets the rule of rescue. *JAMA* 1991;**265**:2218–25, 1991.*

Haig TH, Scott DA, Wickett LI. The rational zero point for an illness index with ratio properties. *Med Care* 1986;**24**:113–24.

Haig TH, Scott DA, Stevens GB. Measurement of the discomfort component of illness. *Med Care* 1989;**27**:280–7.

Hall J, Gerard K, Salkeld G, Richardson J. A cost utility analysis of mammography screening in Australia. *Soc Sci Med* 1992;**34**:993–1004.

Hamel MB, Phillips RS, Davis RB, Desbiens N, Connors AF Jr, Teno JM, *et al.* Outcomes and cost-effectiveness of initiating dialysis and continuing aggressive care in seriously ill hospitalized adults. SUPPORT Investigators. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments. *Ann Intern Med* 1997;**127**:195–202.

Handler RM, Hynes LM, Nease RF Jr. Effect of locus of control and consideration of future consequences on time tradeoff utilities for current health. *Q Life Res* 1997;**6**:54–60.

Hatziandreu EJ, Brown RE, Revicki DA, Turner R, Martindale J, Levine S, Siegel JE. Cost utility of maintenance treatment of recurrent depression with sertraline versus episodic treatment with dothiepin. *PharmacoEconomics* 1994;**5**:249–64.

Hayman JA, Fairclough DL, Harris JR, Weeks JC. Patient preferences concerning the trade-off between the risks and benefits of routine radiation therapy after conservative surgery for early-stage breast cancer. *J Clin Oncol* 1997;**15**:1252–60.

Hays RD, Wu AW, Cleary PD, Fleishman J, Sherbourne CD, Crystal S, *et al.* Associations of time tradeoff with health-related quality of life profile measures in HIV disease. *AHSR FHSR Annu Meeting Abstr Book* 1996;**13**:93–4.

Higgins GL. Discovering a patient's values for advance directives. *Human Med* 1993;**9**:52–6.

Hisashige A, Katayama T, Mikasa H. Quality of life of a genetic disease, mucopolysaccharidosis and anticipation of efficacy and efficiency of its screening in Japan. *Annu Meeting Int Soc Technol Assess Health Care* 1996;**12**:53.

Hopman Rock M, Kraaimaat FW, Bijlsma JW. Quality of life in elderly subjects with pain in the hip or knee. *Q Life Res* 1997;**6**:67–76.

Hornberger JC, Redelmeier DA, Petersen J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *J Clin Epidemiol* 1992;**45**:505–12.

Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A; Brown H. Validity of EuroQoL – a generic health status instrument – in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. *Br J Rheumatol* 1994;**33**:655–62.

Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQoL (EQ-5D). *Br J Rheumatol* 1997;**36**:551–9.

Hutton J, Brown RE, Borowitz M, Abrams K, Rothman M, Shakespeare A. A new decision model for cost–utility comparisons of chemotherapy in recurrent metastatic breast cancer. *PharmacoEconomics* 1996;**9**:8–22.

Hyland ME. Antiasthma drugs: quality-of-life rating scales and sensitivity to longitudinal change. *PharmacoEconomics* 1994;**6**:324–9.

Hyland ME. Quality-of-life measures as providers of information on value-for-money of health interventions. Comparison and recommendations for practice. *PharmacoEconomics* 1997;**11**:19–31.

Hyland ME, Sodergren SC. Development of a new type of global quality of life scale, and comparison of performance and preference for 12 global scales. *Q Life Res* 1996;**5**:469–80.

Irvine EJ. Quality of life in inflammatory bowel disease: biases and other factors affecting scores. *Scand J Gastroenterol Suppl* 1995;**208**:136–40.

Jaeschke R, Guyatt GH, Willan A, Cook D, Harper S, Morris J, *et al.* Effect of increasing doses of beta agonists on spirometric parameters, exercise capacity, and quality of life in patients with chronic airflow limitation. *Thorax* 1994;**49**:479–84.

Johannesson M. QALYs, HYEs and individual preferences – a graphical illustration. *Soc Sci Med* 1994;**39**:1623–32.

Johannesson M, Pliskin JS, Weinstein MC. Are healthy-years equivalents an improvement over quality-adjusted life years? *Med Decis Making* 1993;**13**:281–6.

Johannesson M, Pliskin JS, Weinstein MC. A note on QALYs, time tradeoff, and discounting. *Med Decis Making* 1994;**14**:188–93.

Johannesson M, Jonsson B, Karlson G. Outcome measurement in economic evaluation. *Health Econ* 1996;**5**:279–96.*

Johnson ES, Sullivan SD, Mozaffari E, Langley PC, Bodworth NJ. A utility assessment of oral and intravenous ganciclovir for the maintenance treatment of AIDS-related cytomegalovirus retinitis. *PharmacoEconomics* 1996;**10**:623–9.

Jones PW. Measurement of health in asthma and chronic obstructive airways disease. *Pharm Med* 1992;**6**:13–22.

Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979;**47**:263–91.*

Kahneman D, Tversky A. The psychology of preferences. *Sci Am* 1982;**246**:160–73.*

Kaplan RM, Ernst JA. Do category rating scales produce biased preference weights for a health index? *Med Care* 1983;**21**:193–207.

Kaplan RM, Bush JW, Berry CC. Health status index: category rating versus magnitude estimation for measuring levels of well-being. *Med Care* 1979;**17**:501–25.

Kaplan RM, Debon M, Anderson BF. Effects of number of rating scale points upon utilities in a quality of well-being scale. *Med Care* 1991;**29**:1061–4.

Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures. *Q Life Res* 1993;**2**:467–75.

Kavanagh T, Myers MG, Baigrie RS, Mertens DJ, Sawyer P, Shephard RJ. Quality of life and cardiorespiratory function in chronic heart failure: effects of 12 months' aerobic training. *Heart* 1996;**76**:42–9.

Kennedy W, Reinharz D, Tessier G, Contandriopoulos A, Trabut I, Champagne F, *et al.* Cost utility of chemotherapy and best supportive care in non-small cell lung cancer. *PharmacoEconomics* 1995;**8**:316–23.

Kiberd BA, Jindal KK. Screening to prevent renal failure in insulin dependent diabetic patients: an economic evaluation. *BMJ* 1995;**311**:1595–9.

Kind P, Dolan P. The effect of past and present illness experience on the valuations of health states. *Med Care* 1995;**33**:AS255–63.

Kind P, Rosser R, Williams A. A valuation of qualityof life: some psychometric evidence. In: Jones-Lee MW, editor. The value of life and safety. Amsterdam: North-Holland, 1982;159–70.*

Kitai H, Watanabe H, Kubo T, Hisashige A, Sakurai T, Takeda Y. Utility assessment of pregnancy outcomes. *Annu Meeting Int Soc Technol Assess Health Care* 1995;**11**(abstract 67).

Kodish E, Lantos J, Stocking C, Singer PA, Siegler M, Johnson FL. Bone marrow transplantation for sickle cell disease. A study of parents' decisions. *N Engl J Med* 1991;**325**:1349–53.

Krabbe PF, Essink Bot ML, Bonsel GJ. On the equivalence of collectively and individually collected responses: standard-gamble and time-tradeoff judgments of health states. *Med Decis Making* 1996;**16**:120–32.

Krahn MD, Mahoney JE, Eckman MH, Trachtenberg J, Pauker SG, Detsky AS. Screening for prostate cancer. A decision analytic view. *JAMA* 1994;**272**:773–80.

Kreibich DN, Vaz M, Bourne RB, Rorabeck CH, Kim P, Hardie R, *et al.* What is the best way of assessing outcome after total knee replacement? *Clin Orthop* 1996;221–5.

Kreuter M, Sullivan M, Siosteen A. Sexual adjustment after spinal cord injury-comparison of partner experiences in pre- and postinjury relationships. *Paraplegia* 1994;**32**:759–70.

Krumins PE, Fihn SD, Kent DL. Symptom severity and patients' values in the decision to perform a trans-urethral resection of the prostate. *Med Decis Making* 1988;**8**:1–8.

Kuppermann M, Shiboski S, Feeny D, Elkin EP, Washington AE. Can preference scores for discrete states be used to derive preference scores for an entire path of events? An application to prenatal diagnosis. *Med Decis Making* 1997;**17**:42–55.

Kwa VI, Limburg M, de Haan RJ. The role of cognitive impairment in the quality of life after ischaemic stroke. *J Neurol* 1996;**243**:599–604.

Larsen EB, Gerlach J. Subjective experience of treatment, side-effects, mental state and quality of life in chronic schizophrenic out-patients treated with depot neuroleptics. *Acta Psychiatr Scand* 1996;**93**:381–8.

Laupacis A, Wong C, Churchill D. The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. The Canadian Erythropoietin Study Group. *Control Clin Trials* 1991;**12**:168S–79S.

Laupacis A, Bourne R, Rorabeck C, Feeny D, Wong C, Tugwell P, *et al.* The effect of elective total hip replacement on health-related quality of life. *J Bone Joint Surg Am* 1993;**75**:1619–26.

Laupacis A, Muirhead N, Keown P, Wong C. A disease-specific questionnaire for assessing quality of life in patients on hemodialysis. *Nephron* 1992;**60**:302–6 (erratum: *Nephron* 1992;**61**(2):248).

Laupacis A, Keown P, Pus N, Krueger H, Ferguson B, Wong C, *et al.* A study of the quality of life and cost–utility of renal transplantation. *Kidney Int* 1996;**50**:235–42.

Lawrence WF, Fryback DG, Martin PA, Klein R, Klein BE. Health status and hypertension: a population-based study. *J Clin Epidemiol* 1996;**49**:1239–45.

Lenert LA, Morss S, Goldstein MK, Bergen MR, Faustman WO, Garber AM. Measurement of the validity of utility elicitations performed by computerized interview. *Med Care* 1997;**35**:915–20.

Lenert LA, Soetikno RM. Automated computer interviews to elicit utilities: potential applications in the treatment of deep venous thrombosis. *J Am Med Informatics Assoc* 1997;**4**:49–56.

Lipscomb J. Value preferences for health: meaning, measurement, and use in program evaluation. In: Kane RL, Kane RA, editors. Values and long term care. Lexington, Mass: Lexington Books, 1982;27–83.*

Llewellyn Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients' values in medicine. *Med Decis Making* 1982;**2**:449–62.

Llewellyn Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. Describing health states. Methodologic issues in obtaining values for health states. *Med Care* 1984;**22**:543–52.

Llewellyn Thomas HA, Thiel EC, McGreal MJ. Cancer patients' evaluations of their current health states: the influences of expectations, comparisons, actual health status, and mood. *Med Decis Making* 1992;**12**:115–22.

Llewellyn Thomas HA, Sutherland HJ, Thiel EC. Do patients' evaluations of a future health state change when they actually enter that state? *Med Care* 1993;**31**:1002–12.

Llewellyn Thomas HA, Williams JI, Levy L, Naylor CD. Using a trade-off technique to assess patients' treatment preferences for benign prostatic hyperplasia. *Med Decis Making* 1996;**16**:262–82.

Loomes G. Disparities between health state measures: is there a rational explanation? In: Gerrard W, editor. The economics of rationality. London: Routledge, 1993.*

Loomes G, McKenzie L. The use of QALYs in health care decison making. *Soc Sci Med* 1989;**28**:299–308.*

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. Oxford: OUP, 1996.

McLeod RS, Churchill DN, Lock AM, Vanderburgh S, Cohen Z. Quality of life of patients with ulcerative colitis preoperatively and postoperatively. *Gastroenterol* 1991;**101**:1307–13.

McLeod RS, Taylor BR, O'Connor BI, Greenberg GR, Jeejeebhoy KN, Royall D, *et al.* Quality of life, nutritional status, and gastrointestinal hormone profile following the Whipple procedure. *Am J Surg* 1996;**169**:179–85.

McNeil BJ, Pauker SG, Sox HC, Tversky A. On the elicitation of preferences for alternative therapies. *N Engl J Med* 1982;**306**(21):1259–62.*

Mangione CM, Marcantonio ER, Goldman L, Cook EF, Donaldson MC, Sugarbaker DJ, *et al.* Influence of age on measurement of health status in patients undergoing elective surgery. *J Am Geriatr Soc* 1993;**41**:377–83.

Manzetti JD. Exercise and quality of life in lung transplant candidates. University of Pittsburgh, 1992.

Markides KS, Lee DJ, Ray LA, Black SA. Physicians' ratings of health in middle and old age: a cautionary note. *J Gerontol* 1993;**48**:S24–7.

Marsden FW, Swanson CE. Outcomes after multi-modality treatment of musculoskeletal tumours. *Acta Orthop Scand Suppl* 1997;**273**:101–5.

Matchar DB, McCrory DC, Bennett CL. Treatment considerations for persons with metastatic prostate cancer: survival versus out-of-pocket costs. *Urology* 1997;**49**:218–24.

Mehrez A, Gafni A. Evaluating health related quality of life: an indifference curve interpretation for the time trade-off technique. *Soc Sci Med* 1990;**31**:1281–3.

Mehrez A, Gafni A. The healthy-years equivalents: how to measure them using the standard gamble approach. *Med Decis Making* 1991;**11**:140–6.

Mehrez A, Gafni A. Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. *Med Decis Making* 1993;**13**:287–92.

Miller MD, Ferris DG. Measurement of subjective phenomena in primary care research: the visual analogue scale. *Fam Pract Res J* 1993;**13**:15–24.

Miyamoto JM, Eraker SA. Parameter estimates for a QALY utility model. *Med Decis Making* 1985;**5**:191–213.*

Mohide EA, Torrance GW, Streiner DL, Pringle DM, Gilbert R. Measuring the wellbeing of family caregivers using the time trade-off technique. *J Clin Epidemiol* 1988;**41**:475–82.

Molzahn AE. Perceptions of patients, physicians, and nurses regarding the quality of life of individuals with end stage renal disease. University of Alberta, 1989.

Molzahn AE, Northcott HC, Hayduk L. Quality of life of patients with end stage renal disease: a structural equation model. *Q Life Res* 1996;**5**:426–32.

Molzahn AE, Burton JR, McCormick P, Modry DL, Soetaert P, Taylor P. Quality of life of candidates for and recipients of heart transplants. *Can J Cardiol* 1997;**13**:141–6.

Molzahn AE, Northcott HC, Dossetor JB. Quality of life of individuals with end stage renal disease: perceptions of patients, nurses, and physicians. *ANNA J* 1997;**24**:325–33.

Morss SE, Lenert LA, Faustman WO. The side effects of antipsychotic drugs and patients' quality of life: patient education and preference assessment with computers and multimedia. *Proc Annu Symp Comput Appl Med Care* 1993;17–21.

Murray CJ, Lopez AD. Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: global Burden of Disease Study. *Lancet* 1997;**349**:1347–52.

Mushlin AI, Mooney C, Grow V, Phelps CE. The value of diagnostic information to patients with suspected multiple sclerosis. Rochester–Toronto MRI Study Group. *Arch Neurol* 1994;**51**:67–72.

Nease RF, Owens DK. A method for estimating the cost-effectiveness of incorporating patient preferences into practice guidelines. *Med Decis Making* 1994;**14**:382–92.

Nease R, Hynes L, Littenberg B, Tosteson A, Sumner W, Owens D. Variation in patient preferences for outcomes associated with the management of mild hypertension: implications for practice guidelines. *AHSR FHSR Annu Meeting Abstr Book* 1994;**11**:25.

Nease RF Jr, Kneeland T, O'Connor GT, Sumner W, Lumpkins C, Shaw L, *et al.* Variation in patient utilities for outcomes of the management of chronic stable angina. Implications for clinical practice guidelines. Ischemic Heart Disease Patient Outcomes Research Team. *JAMA* 1995;**273**:1185–90 (erratum: *JAMA* 1995;**274**(8):612).

Nease RF Jr, Tsai R, Hynes LM, Littenberg B. Automated utility assessment of global health. *Q Life Res* 1996;**5**:175–82.

Newbold D. A brief description of the methods of economic appraisal and the valuation of health states. *J Adv Nurs* 1995;**21**:325–33.

Nichol G, Llewellyn Thomas HA, Thiel EC, Naylor CD. The relationship between cardiac functional capacity and patients' symptom-specific utilities for angina: some findings and methodologic lessons. *Med Decis Making* 1996;**16**:78–85.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Planning Manage* 1991;**6**:234–42.

Nord E. EuroQoL: health-related quality of life measurement. Valuations of health states by the general public in Norway. *Health Policy* 1991;**18**:25–36.

Nord E. An alternative to QALYs: the saved young life equivalent (SAVE). *BMJ* 1992;**305**:875–7.*

Nord E. Methods for quality adjustment of life years. *Soc Sci Med* 1992;**34**:559–69.

Nord E. The QALY – a measure of social value rather than individual utility? *Health Econ* 1994;**3**:89–93.*

Nord E. The person-trade-off approach to valuing health care programs. *Med Decis Making* 1995;**15**:201–8.

Nord E. The trade-off between severity of illness and treatment effect in cost-value analysis of health care. *Health Policy* 1993;**24**:227–38.

Nord E, Richardson J, Macarounas Kirchmann K. Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. *Int J Technol Assess Health Care* 1993;**9**:463–78.

Norum J, Wist E. Psychological distress in survivors of Hodgkin's disease. *Support Care Cancer* 1996;**4**:191–5.

O'Brien B, Viramontes JL. Willingness to pay: a valid and reliable measure of health state preference? *Med Decis Making* 1994;**14**:289–97.

O'Brien BJ, Elswood J, Calin A. Willingness to accept risk in the treatment of rheumatic disease. *J Epidemiol Commun Health* 1990;**44**:249–52.

O'Leary JF, Fairclough DL, Jankowski MK, Weeks JC. Comparison of time-tradeoff utilities and rating scale values of cancer patients and their relatives: evidence for a possible plateau relationship. *Med Decis Making* 1995;**15**:132–7.

Oldridge N, Guyatt G, Jones N, Crowe J, Singer J, Feeny D, *et al.* Effects on quality of life with comprehensive rehabilitation after acute myocardial infarction. *Am J Cardiol* 1991;**67**:1084–9.

Oldridge N, Furlong W, Feeny D, Torrance G, Guyatt G, Crowe J, *et al.* Economic evaluation of cardiac rehabilitation soon after acute myocardial infarction. *Am J Cardiol* 1993;**7**2:154–61.

Olsen JA. Persons vs years: two ways of eliciting implicit weights. *Health Econ* 1994;**3**:39–46.

Osoba D. Measuring the effect of cancer on health-related quality of life. *PharmacoEconomics* 1995;**7**:308–19.

Owens DK, Cardinalli AB, Nease RF Jr. Physicians' assessments of the utility of health states associated with human immunodeficiency virus (HIV) and hepatitis B virus (HBV) infection. *Q Life Res* 1997;**6**:77–86.

Padilla GV, Grant MM, Lipsett J, Anderson PR, Rhiner M, Bogen C. Health quality of life and colorectal cancer. *Cancer* 1992;**70**:1450–6.

Patrick DL. Constructing social metrics for health status indexes. *Int J Health Serv* 1976;**6**:443–53.

Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res* 1973;**8**:228–45.*

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Med Decis Making* 1994;**14**:9–18.

Perez DJ, McGee R, Campbell AV, Christensen EA, Williams S. A comparison of time trade-off and quality of life measures in patients with advanced cancer. *Q Life Res* 1997;**6**:133–8.

Pfennings L, Cohen L, van der Ploeg H. Preconditions for sensitivity in measuring change: visual analogue scales compared to rating scales in a Likert format. *Psychol Rep* 1995;**77**:475–80.

Pinto Prades JL. Is the person trade-off a valid method for allocating health care resources? *Health Econ* 1997;**6**:71–81.

Prose R, Conns BB. Lack of benefit of combination chemotherapy for metastatic carcinoma of the great toe. *J Clin Oncol* 1989;**7**:974–8.

Provenzale D, Shearin M, Phillips Bute BG, Drossman DA, Li Z, Tillinger W, *et al.* Health-related quality of life after ileoanal pull-through evaluation and assessment of new health status measures. *Gastroenterology* 1997;**113**:7–14.

Rabin R, Rosser RM, Butler C. Impact of diagnosis on utilities assigned to states of illness. *J R Soc Med* 1993;**86**:444–8.

Ramsey SD, Patrick DL, Albert RK, Larson EB, Wood DE, Raghu G. The cost-effectiveness of lung transplantation. A pilot study. University of Washington Medical Center Lung Transplant Study Group. *Chest* 1995;**108**:1594–601.

Ramsey SD, Patrick DL, Lewis S, Albert RK, Raghu G. Improvement in quality of life after lung transplantation: a preliminary study. The University of Washington Medical Center Lung Transplant Study Group. *J Heart Lung Transplant* 1995;**14**:870–7.

Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes. Comparison of assessment methods. *Med Decis Making* 1984;**4**:315–29.

Redelmeier DA, Heller DN. Time preference in medical decision making and cost-effectiveness analysis. *Med Decis Making* 1993;**13**:212–17.

Reed WW, Herbers JE Jr, Noel GL. Cholesterol-lowering therapy: what patients expect in return. *J Gen Intern Med* 1993;**8**:591–6.

Revicki DA. Relationship between health utility and psychometric health status measures. *Med Care* 1992;**30**:MS274–82.

Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Q Life Res* 1993;**2**:477–87.

Revicki D, Wu A. Transition in HIV disease severity and health status and utility outcomes. *AHSR FHSR Annu Meeting Abstr Book* 1994;**11**:97.

Revicki DA, Wu AW. Discrimination and responsiveness of health status and utility measures. *Annu Meeting Int Soc Technol Assess Health Care* 1995;**11** (abstract 6).

Revicki DA, Wu AW, Murray MI. Change in clinical status, health status, and health utility outcomes in HIV-infected patients. *Med Care* 1995;**33**:AS173–82.

Revicki DA, Shakespeare A, Kind P. Preferences for schizophrenia-related health states: a comparison of patients, caregivers and psychiatrists. *Int Clin Psychopharmacol* 1996;**11**:101–8.

Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994;**39**:7–21.

Rittenhouse BE. Healthy years equivalents versus time trade-off. Ambiguity on certainty and uncertainty. *Int J Technol Assess Health Care* 1997;**13**:35–48.

Robinson A, Dolan P, Williams A. Valuing health states using VAS and TTO: what lies behind the numbers? *Soc Sci Med* 1997;**45**:1289–97.

Robustelli Della Cuna G, Pellegrini A, Piazzi M, Decoster JM, Hearron AE, Silva A, *et al.* Effect of methylprednisolone sodium succinate on quality of life in preterminal cancer patients: a placebo-controlled, multicenter study. *Eur J Cancer Clin Oncol* 1989;**25**:1817–21.

Rosenberg EE, Tannenbaum TN. Measuring health status: an approach for family practice researchers. *Fam Med* 1991;**23**:52–6.

Rosser R, Kind, P. A scale of valuations of states of illness: is there a social consensus? *Int J Epidemiol* 1978;**7**:347–58.

Ruperto N, Ravelli A, Levinson JE, Shear ES, Murray K, Link Tague B, *et al.* Long-term health outcomes and quality of life in American and Italian inception cohorts of patients with juvenile rheumatoid arthritis. II. Early predictors of outcome. *J Rheumatol* 1997;**24**:952–8.

Russel MGV, Pastoor CJ, Brandon S, Rijken J, Engels LGJ, Van der Heijde DMF, Stockbrugger RW. Validation of the dutch translation of the inflammatory dowel disease questionnaire (IBDQ): a health-related duality of life questionnaire in inflammatory bowel disease. *Digestion* 1997;**58**:282–8.

Russell JD, Beecroft ML, Ludwin D, Churchill DN. The quality of life in renal transplantation – a prospective study. *Transplantation* 1992;**54**:656–60.

Rutten van Molken MP, Bakker CH, van Doorslaer EK, van der Linden S. Methodological issues of patient utility measurement. Experience from two clinical trials. *Med Care* 1995;**33**:922–37.

Rutten van Molken MP, Custers F, van Doorslaer EK, Jansen CC, Heurman L, Maesen FP, *et al.* Comparison of performance of four instruments in evaluating the effects of salmeterol on asthma quality of life. *Eur Respir J* 1995;**8**:888–98.

Rutten-van-Molken M, Van-Doorslaer E, Van-den-Boom G. Comparing the responsiveness of descriptive and preference based quality of life measures in asthma. *Annu Meeting Int Soc Technol Assess Health Care* 1996;**12**:6.

Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chronic Dis* 1981;**31**:697–704.*

Sculpher MJ, Dwyer N, Byford S, Stirrat GM. Randomised trial comparing hysterectomy and transcervical endometrial resection: effect on health related quality of life and costs two years after surgery. *Br J Obstet Gynaecol* 1996;**103**:142–9.

Selby P. Measurement of the quality of life after cancer treatment. *Br J Hosp Med* 1985;**33**:266–71.

Sesso R, Yoshihiro MM, Ajzen H. Late diagnosis of chronic renal failure and the quality of life during dialysis treatment. *Braz J Med Biol Res* 1996;**29**:1283–9.

Shackley P, Cairns J. Evaluating the benefits of antenatal screening: an alternative approach. *Health Policy* 1996;**36**:103–15.

Shaul MP. The examination of 3 rating-scales measuring perceptions of quality-of-life, well-being and the impact of rheumatoid-arthritis (RA) on everyday life. *Arthritis Rheumat* 1995;**38**:1405.

Simpson KN. Design and measurement issues in economic studies of costly chronic illnesses: examples from HIV-drug therapy. *Annu Meeting Int Soc Technol Assess Health Care* 1995;**11**(abstract 83).

Sintonen H. An approach to measuring and valuing health states. *Soc Sci Med* 1981;**15C**:55–65.*

Sivertssen E, Field NB, Abdelnoor M. Quality of life after open heart surgery. *Vasc Surg* 1994;**28**:581–8.

Smith RD, Hall J, Gurney H, Harnett PR. A cost–utility approach to the use of 5-fluorouracil and levamisole as adjuvant chemotherapy for Dukes' C colonic carcinoma. *Med J Aust* 1993;**158**:319–22.

Stalmeier PF, Bezembinder TG, Unic IJ. Proportional heuristics in time tradeoff and conjoint measurement. *Med Decis Making* 1996;**16**:36–44.

Stevens SS. Issues in psychophysical measurement. *Psychol Rev* 1971;**78**:426–50.*

Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Med Decis Making* 1994;**14**:82–90.

Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Habbema JD, de Haes JC. The "utility" of the time trade-off method in cancer patients: feasibility and proportional trade-off. *J Clin Epidemiol* 1995;**48**:1207–14.

Stiggelbout AM, de Haes JC, Kiebert GM, Kievit J, Leer JW. Tradeoffs between quality and quantity of life: development of the QQ Questionnaire for Cancer Patient Attitudes. *Med Decis Making* 1996;**16**:184–92.

Stiggelbout AM, Eijkemans MJ, Kiebert GM, Kievit J, Leer JW, De Haes HJ. The 'utility' of the visual analog scale in medical decision making and technology assessment. Is it an alternative to the time trade-off? *Int J Technol Assess Health Care* 1996;**12**:291–8.

Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. *Phys Ther* 1996;**76**:1109–23.

Sullivan M, Katon WJ, Russo J, Dobie R, Sakai C. Somatization, co-morbidity, and the quality of life: measuring the effect of depression upon chronic medical illness. *Psychiatr Med* 1992;**10**:61–76.

Sumner W, Nease R, Littenberg B. U-titer: a utility assessment tool. *Proc Annu Symp Comput Appl Med Care* 1991;701–5.

Sutherland HJ, Llewellyn Thomas H, Boyd D, Till JE. Attitudes towards quality of survival: the concept of maximum endurable time. *Med Decis Making* 1982;**2**:299–309.*

Sutherland HJ, Dunn V, Boyd NF. Measurement of values for states of health with linear analog scales. *Med Decis Making* 1983;**3**:477–87.

Swan JS, Fryback DG, Lawrence WF, Katz DA, Helsey DM, Hagenauer ME, *et al.* MR and conventional angiography: work in progress toward assessing utility in radiology. *Acad Radiol* 1997;**4**:475–82.

Tamburini M, Filiberti A, Barbieri A, Zanoni F, Pizzocaro G, Barletta L, *et al.* Psychological aspects of testis cancer therapy: a prospective study. *J Urol* 1989;**142**:1487–9.

Thompson MS. Willingness to pay to accept risks to cure chronic disease. *Am J Public Health* 1986;**76**:392–6.*

Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Econ Plan Sci* 1976;**10**:129–36.*

Torrance GW. Preferences for health states: a review of measurement methods. *Mead Johnson Symp Perinat Dev Med* 1982;37–45.

Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986;**5**:1–30.

Torrance GW. Utility approach to measuring health-related quality of life. *J Chronic Dis* 1987;**40**:593–603.

Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol Assess Health Care* 1989;**5**:559–75.*

Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care* 1996;**34**:702–22.

Tousignant P, Cosio MG, Levy RD, Groome PA. Quality adjusted life years added by treatment of obstructive sleep apnea. *Sleep* 1994;**17**:52–60.

Tsevat J. Methods for assessing health-related quality of life in HIV-infected patients. *Psychol Health* 1994;**9**:19–30.

Tsevat J, Goldman L, Soukup JR, Lamas GA, Connors KF, Chapin CC, Lee TH. Stability of time-tradeoff utilities in survivors of myocardial infarction. *Med Decis Making* 1993;**13**:161–5.

Tsevat J, Cook EF, Green ML, Matchar DB, Dawson NV, Broste SK, *et al.* Health values of the seriously ill. SUPPORT investigators. *Ann Intern Med* 1995;**122**:514–20.

Tsevat J, Solzan JG, Kuntz KM, Ragland J, Currier JS, Sell RL, Weinstein MC. Health values of patients infected with human immunodeficiency virus. Relationship to mental health and physical functioning. *Med Care* 1996;**34**:44–57.

Tversky A, Slovic P, Kahneman D. The causes of preference reversal. *Am Econ Rev* 1990;**80**:204–17.*

Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Individual utilities are inconsistent with rationing choices: a partial explanation of why Oregon's cost-effectiveness list failed. *Med Decis Making* 196;**16**:108–16.

Ure BM, Slany E, Eypasch EP, Gharib M, Holschneider AM, Troidl H. Long-term functional results and quality of life after colon interposition for long-gap oesophageal atresia. *Eur J Pediatr Surg* 1995;**5**:206–10.

Van Agt HME, Essink Bot M, Krabbe PFM, Bonsel GJ. Retest reliability of health state valuations collected with the EuroQoL questionnaire. *Soc Sci Med* 1994;**39**:1537–44.

van-Busschbach JJ, Hessing DJ, de-Charro FT. Assessing the quality of life: an empirical comparison of four different methods of assessment. *Abstr Int Soc Technol Assess Health Care* 1992;18.

van-Busschbach J, de-Wit A, de-Charro F. Growth hormone treatment of patients with a short stature: costs and effects on the quality of life. *Abstr Int Soc Technol Assess Health Care* 1993;**9**:165.

Veit CT, Rose BJ, Ware JE Jr. Effects of physical and mental health on health-state preferences. *Med Care* 1982;**20**:386–401.

Viramontes JL, O'Brien BJ. Empirical evidence on the validity and reliability of willingness-to-pay (WTP) as a measure of health state preference. *Abstr Int Soc Technol Assess Health Care* 1993;**9**:61.

Vollmer Conna U, Hickie I, Hadzi Pavlovic D, Tymms K, Wakefield D, Dwyer J, Lloyd A. Intravenous immunoglobulin is ineffective in the treatment of patients with chronic fatigue syndrome. *Am J Med* 1997;**103**:38–43.

Wakker P, Stiggelbout A. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Med Decis Making* 1995;**15**:180–6.

Weeks J. Taking quality of life into account in health economic analyses. *J Natl Cancer Inst Monogr* 1996;23–7.

Wilkinson TJ, Hanger HC, Elmslie J, George PM, Sainsbury R. The response to treatment of subclinical thiamine deficiency in the elderly. *Am J Clin Nutr* 1997;**66**:925–8.

Williams A. EuroQoL – a new facility for the measurement of health-related quality of life. *Health Policy* 1990;**16**:199–208.

Wolfe F, Hawley DJ, Cathey MA. Measurement of gold treatment effect in clinical practice: evidence for effectiveness of intramuscular gold therapy. *J Rheumatol* 1993;**20**:797–802.

Wolfe F. Health status questionnaires. *Rheum Dis Clin North Am* 1995;**21**:445–64.

Wolfson AD, Sinclair AJ, Bombardier C, McGreer A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA, editors. Values and long term care. Lexington: Lexington Books, 1982;191–214.*

Zathraeus N, Johannesson M, Henriksson P, Strand RT. The impact of hormone replacement therapy on quality of life and willingness to pay. *Br J Obstet Gynaecol* 1997;**104**:1191–5.

Ziegler F. Cost–utility in CNS drug trials. *Eur Psychiat* 1996;**11**:159–64.

Zug KA, Littenberg B, Baughman RD, Kneeland T, Nease RF, Sumner W, *et al.* Assessing the preferences of patients with psoriasis. A quantitative, utility approach. *Arch Dermatol* 1995;**131**:561–8.

Appendix 3

Supplementary data for chapter 5

Papers on five MAUSs identified by search

Quality of Well-Being Scale

Anderson GM. A comment on the index of well-being. *Med Care* 1982;**20**:513–15.

Anderson JP, Bush JW, Berry CC. Classifying function for health outcome and quality-of-life evaluation. Self- versus interviewer modes. *Med Care* 1986;**24**:454–69.

Anderson JP, Bush JW, Berry CC. Internal Consistency Analysis: a method for studying the accuracy of function assessment for health outcome and quality of life evaluation. *J Clin Epidemiol* 1988;**41**:127–37.

Anderson JP, Kaplan RM, Berry CC, Bush JW, Rumbaut RG. Interday reliability of function assessment for a health-status measure – the quality of well-being scale. *Med Care* 1989;**27**:1076–84.

Anderson JP, Kaplan RM, Schneiderman LJ. Effects of offering advance directives on quality adjusted life expectancy and psychological well-being among ill adults. *J Clin Epidemiol* 1994;**47**:761–72.

Andresen EM, Patrick DL, Carter WB, Malmgren JA. Comparing the performance of health status measures for healthy older adults. *J Am Geriatr Soc* 1995;**43**:1030–4.

Bakker CH, Rutten van Molken M, van Doorslaer E, Bennett K, van der Linden S. Health related utility measurement in rheumatology: an introduction. *Patient Educ Couns* 1993;**20**:145–52.

Balaban DJ, Sagi PC, Goldfarb NI, Nettler S. Weights for scoring the quality of well-being instrument among rheumatoid arthritics. A comparison to general population weights. *Med Care* 1986;**24**:973–80.

Bombardier C, Raboud J. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. The Auranofin Cooperating Group. *Control Clin Trials* 1991;**12**:243S–56S.

Bombardier C, Ware J, Russell I, Larson MG, Chalmers A, Leighton Read J. Auranofin therapy and quality of life in patients with rheumatoid arthritis. *Am J Med* 1986;**81**:565–78.

Bradlyn AS, Harris CV, Warner JE, Ritchey AK, Zaboy K. An investigation of the validity of the quality of Well-Being Scale with pediatric oncology patients. *Health Psychol* 1993;**12**:246–50.

Bush JW, Anderson JP, Kaplan RM, Blischke WR. Counter-intuitive preferences in health-related quality-of-life measurement. *Med Care* 1982;**20**:516–25.

Calfas KJ, Kaplan RM, Ingram RE. One-year evaluation of cognitive-behavioral intervention in osteoarthritis. *Arthritis Care Res* 1992;**5**:202–9.

de Groot J, de Groot W, Kamphuis M, Vos PF, Berend K, Blankestijn PJ. Kwaliteit van leven van dialysepatienten in Utrecht en Willemstad weinig verschillend [Little difference in quality of life of dialysis patients in Utrecht and Willemstad]. *Ned Tijdschr Geneeskd* 1994;**138**:862–6.

Dirksen SR. Search for meaning in long-term cancer survivors. *J Adv Nurs* 1995;**21**:628–33.

Elvik R. The validity of using health state indexes in measuring the consequences of traffic injury for public-health. *Soc Sci Med* 1995;**40**:1385–98.

Erickson P, Kendall EA, Anderson JP, Kaplan RM. Using composite health status measures to assess the nation's health. *Med Care* 1989;**27**:S66–76.

Fryback DG, Dasbach ED, Klein R, Klein BEK, Martin PA, Dorn N, *et al.* Health assessment by SF-36, Quality of Well-Being Index and time trade-offs: predicting one measure from another. *Med Decis Making* 1992;**12**:348P.

Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K, Martin P.A. The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors. *Med Decis Making* 1993;**13**:89–102.

Ganiats TG, Palinkas LA, Kaplan RM. Comparison of Quality of Well-Being Scale and Functional Status Index in patients with atrial fibrillation. *Med Care* 1992;**30**:958–64.

Gilbert A, Owen N, Innes JM, Sansom L. Trial of an intervention to reduce chronic benzodiazepine use among residents of aged-care accommodation. *Aust NZ J Med* 1993;**23**:343–7.

Holbrook TL, Hoyt DB, Anderson JP, Hollingsworth-Fridlund P, Shackford SR. Functional limitation after major trauma: a more sensitive assessment using the Quality of Well-Being Scale – the trauma recovery pilot project. *J Trauma* 1994;**36**:74–8.

Hornberger JC, Redelmeier DA, Petersen J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *J Clin Epidemiol* 1992;**45**:505–12.

Kaplan RI, Atkins CJ. The well-year of life as a basis for patient decision-making. *Patient Educ Couns* 1989;**13**:281–95.

Kaplan RM. Health outcome models for policy analysis. *Health Psychol* 1989;**8**:723–35.

Kaplan RM. Application of a general health policy model in the American health care crisis. *J R Soc Med* 1993;**86**:277–81.

Kaplan RM. Quality of life assessment for cost/utility studies in cancer. *Cancer Treat Rev* 1993; **19**(Suppl A): 85–96.

Kaplan RM. Value judgment in the oregon medicaid experiment. *Med Care* 1994;**32**:975–88.

Kaplan RM. Using quality-of-life information to set priorities in health-policy. *Social Indicators Res* 1994;**33**:121–63.

Kaplan RM, Anderson JP. A general health policy model: update and application. *Health Services Res* 1988;**23**:203–35.

Kaplan RM, Bush JW. Health-related quality of life measurement for evaluation research and policy analysis. *Health Psychol* 1982;**1**:61–80.

Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. *Health Serv Res* 1976;**11**:478–507.

Kaplan RM, Bush JW, Berry CC. Health status index: category rating versus magnitude estimation for measuring levels of well-being. *Med Care* 1979;**17**:501–25.

Kaplan RM, Atkins CJ, Timms, R. Validity of a quality of well-being scale as an outcome measure in chronic obstructive pulmonary disease. *J Chronic Dis* 1984;**37**:85–95.

Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F, Orenstein D. The Quality of Well-being Scale. Applications in AIDS, cystic fibrosis, and arthritis. *Med Care* 1989;**27**:S27–43.

Kaplan RM, Anderson JP, Wingard DL. Gender differences in health-related quality of life. *Health Psychol* 1991;**10**:86–93.

Kaplan RM, Debon M, Anderson BF. Effects of number of rating scale points upon utilities in a Quality of Well-Being Scale. *Med Care* 1991;**29**:1061–4.

Kaplan RM, Coons SJ, Anderson JP. Quality of life and policy analysis in arthritis. *Arthritis Care Res* 1992;**5**:173–83.

Kaplan RM, Anderson JP, Patterson TL, McCutchan JA, Weinrich JD, Heaton RK, *et al.* Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Center. *Psychosom Med* 1995;**57**:138–47.

Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;**28**:632–42.

Manzetti JD, Hoffman LA, Sereika SM, Sciurba FC, Griffith BP. Exercise, education, and quality of life in lung transplant candidates. *J Heart Lung Transplant* 1994;**13**:297–305.

Mold JW, Holtgrave DR, Bisonni RS, Marley DS, Wright RA, Spann SJ. The evaluation and treatment of men with asymptomatic prostate nodules in primary care: a decision analysis. *J Fam Pract* 1992;**34**:561–8.

Nord E. Unjustified use of the Quality of Well-Being Scale in priority setting in Oregon. *Health Policy* 1993;**24**:45–53.

Orenstein DM, Kaplan RM. Measuring the quality of well-being in cystic fibrosis and lung transplantation. The importance of the area under the curve. *Chest* 1991;**100**:1016–18.

Orenstein DM, Nixon PA, Ross EA, Kaplan RM. The quality of well-being in cystic fibrosis. *Chest* 1989;**95**:344–7.

Orenstein DM, Pattishall EN, Nixon PA, Ross EA, Kaplan RM. Quality of well-being before and after antibiotic treatment of pulmonary exacerbation in patients with cystic fibrosis. *Chest* 1990;**98**:1081–4.

Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res* 1973;**8**:228–45.

Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. *J Health Soc Behav* 1973;**14**:6–23.

Read JL, Quinn RJ, Hoefer MA. Measuring overall health: an evaluation of three important approaches. *J Chronic Dis* 1987;**40**(Suppl 1):7S–26S.

Reed PG. Religiousness among terminally ill and healthy adults. *Res Nurs Health* 1986;**9**:35–41.

Schneiderman LJ, Kronick R, Kaplan RM, Anderson JP, Langer RD. Effects of offering advance directives on medical treatments and costs. *Ann Intern Med* 1992;**117**:599–606.

Tandon PK, Stander H, Schwarz RP Jr. Analysis of quality of life data from a randomized, placebo-controlled heart-failure trial. *J Clin Epidemiol* 1989;**42**:955–62.

Tramarin A, Milocchi F, Tolley K, Vaglia A, Marcolini F, Manfrin V, *et al.* An economic evaluation of home-care assistance for AIDS patients: a pilot study in a town in northern Italy. *Aids* 1992;**6**:1377–83.

Visser MC, Fletcher AE, Parr G, Simpson A, Bulpitt CJ. 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. *J Clin Epidemiol* 1994;**47**:157–63.

Wu AW, Mathews WC, Brysk LT, Hampton Atkinson J, Grant I, Abramson I, *et al.* DD. Quality of life in a placebo-controlled trial of Zidovudine in patients with AIDS and AIDS-related complex. *J AIDS* 1990;**3**:683–90.

Rosser disability and distress scale

Bryan S, Parkin D, Donaldson C. Chiropody and the QALY – a case-study in assigning categories of disability and distress to patients. *Health Policy* 1991;**18**:169–85.

Carr-Hill RA, Morris J. Current practice in obtaining the "Q" in QALYs: a cautionary note. *BMJ* 1991;**303**:699–701.

Chan CLH, Villar RN. Obesity and quality-of-life after primary hip-arthroplasty. *J Bone Joint Surg Br Vol* 1996;**78B**:78–81.

Coast J. Reprocessing data to form QALYs. *BMJ* 1992;**305**:87–90.

Cole RP, Shakespeare V, Shakespeare P, Hobby JA. Measuring outcome in low-priority plastic surgery patients using quality of life indices. *Br J Plast Surg* 1994;**47**:117–21.

Donaldson C, Atkinson A, Bond J, Wright K. QALYS and long-term care for elderly people in the UK: scales for assessment of quality of life. *Age Ageing* 1988;**17**:379–87.

Donaldson C, Atkinson A, Bond J, Wright K. Should QALYs be programme-specific? *J Health Econ* 1988;**7**:239–57.

Elvik R. The validity of using health state indexes in measuring the consequences of traffic injury for public-health. *Soc Sci Med* 1995;**40**:1385–98.

Gater RA, Kind P, Gudex C. Quality of life in liaison psychiatry. A comparison of patient and clinician assessment. *Br J Psychiat* 1995;**166**:515–20.

Glasziou PP, Bromwich S, Simes RJ. Quality of life six months after myocardial infarction treated with thrombolytic therapy. AUS-TASK Group. Australian arm of International tPA/SK Mortality Trial. *Med J Aust* 1994;**161**:532–6.

Gudex C. QALYs and their use by the health service. Discussion paper 20. York: Centre for Health Economics, University of York, 1986.

Gudex CM. Health-related quality of life in endstage renal failure. *Q Life Res* 1995;**4**:359–66.

Gudex C, Kind P. The QALY toolkit. Centre for Health Economics discussion paper 93. York: Centre for Health Economics, University of York, 1988.

Gudex C, Kind P. Chiropody and the QALY – a case-study in assigning categories and distress to patients. *Health Policy* 1991;**19**:79–80.

Gudex C, Williams A, Jourdan M, Mason R, Maynard J, O'Flynn R, Rendall M. Prioritising waiting lists. *Health Trends* 1990;**22**:103–8.

Hollingworth W, Mackenzie R, Todd CJ, Dixon AK. Measuring changes in quality-of-life following magnetic-resonance- imaging of the knee – SF-36, EuroQoL((c)) or Rosser index. *Q Life Res* 1995;**4**:325–34.

Humphreys WV, Evans F, Watkin G, Williams T. Critical limb ischemia in patients over 80 years of age – options in a district general-hospital. *Br J Surg* 1995;**82**:1361–3.

Kallis P, Unsworth White J, Munsch C, Gallivan S, Smith EE, Parker DJ, *et al.* Disability and distress following cardiac surgery in patients over 70 years of age. *Eur J Cardiothorac Surg* 1993;**7**:306–11.

Kind P. Measuring valuations for health states: a survey of patients in general practice. Centre for Health Economics discussion paper 76. York: Centre for Health Economics, University of York, 1990.

Kind P, Gudex C. The HMQ: measuring health status in the community. Centre for Health Economics discussion paper 93. York: Centre for Health Economics, University of York, 1991.

Kind P, Gudex CM. Measuring health-status in the community – a comparison of methods. *J Epidemiol Commun Health* 1994;**48**:86–91.

Kind P, Rosser R. The quantification of health. *Eur J Social Psychol* 1988;**18**:63–77.

Kind P, van Dalen H, Morris J, Williams A. Comparing saling methods: Rosser revisited. Centre for Health Economics discussion paper 107. York: Centre for Health Economics, University of York, 1993.

Launois R, Henry B, Marty JR, Gersberg M, Lassale C, Benoist M, Goehrs JM. Chemonucleolysis versus surgical diskectomy for sciatica secondary to lumbar disc herniation – a cost and quality-of-life evaluation. *PharmacoEconomics* 1994;**6**:453–63.

Lonnqvist J, Sihvo S, Syvalahti E, Sintonen H, Kiviruusu O, Pitkanen H. Moclobemide and fluoxetine in the prevention of relapses following acute treatment of depression. *Acta Psychiatr Scand* 1995;**91**:189–94.

Mackenzie R, Hollingworth W, Dixon AK. Quality of life assessments in the evaluation of magnetic resonance imaging. *Q Life Res* 1994;**3**:29–37.

Magee TR, Scott DJ, Dunkley A, St Johnston J, Campbell WB, Baird RN, *et al.* Quality of life following surgery for abdominal aortic aneurysm. *Br J Surg* 1992;**79**:1014–16.

Normantaylor FH, Palmer CR, Villar RN. Quality-of-life improvement compared after hip and knee replacement. *J Bone Joint Surg Br Vol* 1996;**78B**:74–7.

Payne SP, Galland RB. The use of a simple clinical cardiac risk index predictive of long-term outcome after infrarenal aortic reconstruction. *Eur J Vasc Endovasc Surg* 1995;**9**:138–42.

Petrou S, Davey P, Malek M. The application of the Rosser–Kind classification to hip and knee joint replacement surgery. paper presented at Health Economists Study Group meeting Sheffield, 1992.

Rabin R, Rosser RM, Butler C. Impact of diagnosis on utilities assigned to states of illness. *J R Soc Med* 1993;**86**:444–8.

Rawles J, Light J, Watt M. Quality of life in the first 100 days after suspected acute myocardial infarction – a suitable trial endpoint? *J Epidemiol Commun Health* 1992;**46**:612–16.

Rosser RM, Kind P. A scale of valuations of states of illness: is there a social consensus? *Int J Epidemiol* 1978;**7**:347–58.

Rosser RM, Watts VC. The measurement of hospital output. *Int J Epidemiol* 1972;**1**:361–8.

Rosser R, Allison R, Butler C, Cottee M, Rabin R, Selai C. 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. Lancaster: MTP Press, 1993;179–84.

Unsworthwhite J, Kallis P, Treasure T, Pepper JR. Quality-of-life after cardiac-surgery in patients over 70 years of age. *Cardiol Elderly* 1994;**2**:133–8.

van Dalen H, Williams A, Gudex C. Lay peoples evaluations of health – are there variations between different subgroups. *J Epidemiol Commun Health* 1994;**48**:248–53.

Wade DT. The Q in QALYs. *BMJ* 1991;**303**:1136–7.

Watkins LD, Bell BA, Marsh HT, Uttley D. A scale for neurosurgical audit. *Br J Neurosurg* 1990;**4**:463–5.

Whynes DK, Neilson AR. Convergent validity of two measures of the quality of life. *Health Econ* 1993;**2**:229–35.

Whynes DK, Neilson AR, Robinson MH, Hardcastle JD. Colorectal cancer screening and quality of life. *Q Life Res* 1994;**3**:191–8.

Williams A. Economics of coronary artery bypass grafting. *BMJ* 1985;**291**:326–9.

HUI-II and HUI-III

Barr RD, Furlong W, Dawson S, Whitton AC, Strautmanis I, Pai M, *et al.* An assessment of global health status in survivors of acute lymphoblastic leukemia in childhood. *Am J Pediatr Hematol Oncol* 1993;**15**:284–90.

Barr RD, Pai MKR, Weitzman S, Feeny D, Furlong W, Rosenbaum P, *et al.* A multi-attribute approach to health status measurement and clinical management – illustrated by an application to brain tumors in childhood. *Int J Oncol* 1994;**4**:639–48.

Barr RD, Feeny D, Furlong W, Weitzman S, Torrance GW. A preference-based approach to health-related quality-of-life for children with cancer. *Int J Pediat Hematol/Oncol* 1995;**2**:305–15.

Boyle MH, Torrance GW. Developing multiattribute health indexes. *Med Care* 1984;**22**:1045–57.

Boyle MH, Torrance GW, Sinclair JC, Horwood SP. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *N Engl J Med* 1983;**308**:1330–7.

Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the Health Utilities Index – Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. *Q Life Res* 1995;**4**:249–57.

Cadman D, Goldsmith C. Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. *J Chron Dis* 1986;**39**:643–51.

Cadman D, Goldsmith C, Bashim P. Values, preferences and decisions in the care of children with developmental disabilities. *Dev Behav Pediat* 1984;**5**:60–4.

de Groot J, de Groot W, Kamphuis M, Vos PF, Berend K, Blankestijn PJ. Kwaliteit van leven van dialysepatienten in Utrecht en Willemstad weinig verschillend [Little difference in quality of life of dialysis patients in Utrecht and Willemstad]. *Ned Tijdschr Geneeskd* 1994;**138**:862–6.

Elvik R. The validity of using health state indexes in measuring the consequences of traffic injury for public-health. *Soc Sci Med* 1995;**40**:1385–98.

Erickson P, Kendall EA, Anderson JP, Kaplan RM. Using composite health status measures to assess the nation's health. *Med Care* 1989;**27**:S66–76.

Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. *J Clin Oncol* 1992;**10**:923–8.

Feeny D, Leiper A, Barr RD, Furlong W, Torrance GW, Rosenbaum P, Weitzman S. The comprehensive assessment of health status in survivors of childhood cancer: application to high-risk acute lymphoblastic leukaemia. *Br J Cancer* 1993;**67**:1047–52.

Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. *PharmacoEconomics* 1995;**7**:490–502.

Furlong W, Torrance GW, Feeny D. Properties of Health Utilities Index: preliminary evidence. *Q Life Newslett* 1995;3–10.

Gold M, Franks P, Erickson P. Assessing the health of the nation: the predictive value of a preference based measure and self-rated health. *Med Care* 1996;**34**:163–77.

Saigal S, Feeny D, Furlong W, Rosenbaum P, Burrows E, Torrance G. 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. *J Pediatr* 1994;**125**:418–25.

Saigal S, Rosenbaum PL, Furlong WJ, Feeny DH, Burrows E. Self-assessment of their own health-status by extremely low-birth-weight and control teenagers using a multi-attribute health-status classification-system. *Pediatric Res* 1995;**37**:A271.

Torrance GW, Boyle MH, Horwood SP. Applications of Multi-Attribute Utility Theory to measure social preferences for health states. *Operations Res* 1982;**30**:1043–69.

Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health Utilities Index. *PharmacoEconomics* 1995;**7**:503–20.

Verhoef CG, Verbeek AL, Stalpers LJ, van Daal WA. Utiliteitsmeting bij de klinische besluitvorming [Utility assessment in clinical decision making]. *Ned Tijdschr Geneeskd* 1990;**134**:2195–200.

15D

Apajasalo M, Sintonen H, Holmberg C, Sinkkonen J, Aalberg V, Pihko H, *et al.* Quality-of-life in early adolescence – a 16-dimensional health-related measure (16D). *Q Life Res* 1996;**5**:205–11.

Lonnqvist J, Sintonen H, Syvalahti E, Appelberg B, Koskinen T, Mannikko T, *et al.* Antidepressant efficacy and quality of life in depression: a double-blind study with moclobemide and fluoxetine. *Acta Psychiatr Scand* 1994;**89**:363–9.

Lonnqvist J, Sihvo S, Syvalahti E, Sintonen H, Kiviruusu O, Pitkanen H. Moclobemide and fluoxetine in the prevention of relapses following acute treatment of depression. *Acta Psychiatr Scand* 1995;**91**:189–94.

Rissanen P, Aro S, Slatis P, Sintonen H, Paavolainen P. Health and quality of life before and after hip or knee arthroplasty. *J Arthroplasty* 1995;**10**:169–75.

Rissanen P, Aro S, Sintonen H, Slatis P, Paavolainen P. Quality-of-life and functional ability in hip and knee replacements – a prospective-study. *Q Life Res* 1996;**5**:56–64.

Sintonen H. An approach to measuring and valuing health states. *Soc Sci Med* 1981;**15C**:55–65.

Sintonen H. Terveyteen liittyvan elamanlaadun mittaamisesta [Health-related quality of life measures]. *Sairaanhoitaja* 1993;17–19.

EuroQoL (EQ-5D and EQ-6D)

Anderson RT, Aaronson NK, Wilkin D. Critical review of the international assessments of health-related quality of life. *Q Life Res* 1993;**2**:369–95.

Bjork S. EuroQoL conference proceedings. Swedish Health Economics Institute discussion paper 1, 1991.

Brazier J, Jones N, Kind P. Testing the validity of the EuroQoL and comparing it with the SF-36 health survey questionnaire. *Q Life Res* 1993;**2**:169–80.

Brazier J, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and EuroQoL on an elderly population. *Q Life Res* 1996;**5**:195–204.

Brooks RG, Jendteg S, Lindgren B, Persson U, Bjork S. EuroQoL: health-related quality of life measurement. Results of the Swedish questionnaire exercise. *Health Policy* 1991;**18**:37–48.

Caperna J, Mathews WC. 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. *J Invest Med* 1996;**44**:A155.

Carr-Hill RA. A good measure for Eurohealth? *Health Serv J* 1991;**101**:24–5.

Carr-Hill RA. A second opinion: health-related quality of life measurement – Euro style. *Health Policy* 1992;**20**:321–8.

Carr-Hill RA. Health related quality-of-life measurement – Euro style. *Health Policy* 1992;**20**:321–8.

Dolan P. Search for a critical-appraisal of EuroQoL – a response by the EuroQoL group to Gafni and Birch. *Health Policy* 1994;**28**:67–9.

Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQoL: Results from a UK general population survey. Centre for Health Economics discussion paper 138. York: Centre for Health Economics, University of York, 1995.

Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996;**2**:209–32.

Elvik R. The validity of using health state indexes in measuring the consequences of traffic injury for public-health. *Soc Sci Med* 1995;**40**:1385–98.

Essink-Bot ML, Bonsel GJ, Van Der Maas PJ. Valuation of health states by the general public: feasibility of a standardized measurement procedure. *Soc Sci Med* 1990;**31**:1201–6.

Essink-Bot ML, Stouthard ME, Bonsel GJ. Generalizability of valuations on health states collected with the EuroQoLc-questionnaire. *Health Econ* 1993;**2**:237–46.

Essink-Bot ML, Vanroyen L, Krabbe P, Bonsel GJ, Rutten FFH. The impact of migraine on health-status. *Headache* 1995;**35**:200–6.

EuroQoL Group. EuroQoL – a new facility for the measurement of health-related quality-of-life. *Health Policy* 1990;**16**:199–208.

EuroQoL Group. Not a quick fix (response to Carr-Hill). *Health Serv J* 1991;**101**:29.

EuroQoL Group. EuroQoL – a reply and reminder. *Health Policy* 1992;**20**:329–32.

Gravelle H. Valuations of EuroQoL health states: comments and suggestions. Paper presented at the ESRC/SHHD Workshop on Quality of Life, Edinburgh, unpublished, 1995.

Hollingworth W, Mackenzie R, Todd CJ, Dixon AK. Measuring changes in quality-of-life following magnetic-resonance-imaging of the knee – SF-36, EuroQoL((c)) or Rosser index. *Q Life Res* 1995;**4**:325–34.

Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H. Validity of EuroQoL – a generic health status instrument – in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. *Br J Rheumatol* 1994;**33**:655–62.

Kind P. An interim tariff for EuroQoL health states. Personal communication, 1994.

Kind P. The EuroQoL instrument: an index of health - related quality of life. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials, 2nd edn. Philadelphia, PA: Lippincott-Rivera, 1996;191–201.

Kind P, Gudex C, Dolan P, Williams A. Practical and methodological issues in the development of the EuroQoL: the York experience. *Adv Med Sociol* 1994;**5**:219–53.

Kind P, Gudex C, Dolan P; Williams A. Practical and methodological issues in the development of the EuroQoL: the York experience. In: Albrecht GL, Fitzpatrick R, editors. Advances in medical sociology. Greenwich, CT: *J A I* 1994;219–53.

MVH Group. The measurement and valuation of health: first report on the main survey. York: Centre for Health Economics, University of York, 1994.

MVH Group. The measurement and valuation of health: final report on the modelling of valuation tariffs. York: Centre for Health Economics, University of York, 1995.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Plann Manage* 1991;**6**:234–42.

Nord E. EuroQoL – health-related quality-of-life measurement – valuations of health states by the general public in Norway. *Health Policy* 1991;**18**:25–36.

O'Hanlon M, Fox Rushby J, Buxton MJ. A qualitative and quantitative comparison of the EuroQoL and time-trade-off techniques. *Int J Health Serv* 1994;**5**:85–97.

Parkin D. Valuing health states: an exploratory data analysis approach. Paper presented to a meeting of the Health Economists Study Group, University of Oxford, 1991.

Rosser R, Sintonen H. The EuroQoL quality of life project. In: Quality of life assessment: key issues in the 1990s. Lancaster: MTP Press, 1993;197–9.

Sculpher M, Bryan S, Dwyer N, Hutton J, Stirrat GM. An economic evaluation of transcervical endometrial resection versus abdominal hysterectomy for the treatment of menorrhagia. *Br J Obstet Gynaecol* 1993;**100**:244–52.

Selai C, Rosser R. Eliciting EuroQoL descriptive data and utility scale values from inpatients – a feasibility study. *PharmacoEconomics* 1995;**8**:147–58.

Spiegelhalter DJ. The choice of "tariff": comments on the measurement and valuation of health project. University of York, unpublished manuscript, 1995.

Thomas R, Thomson K. Health related quality of life: Technical report. London: SCPR.

van-Agt HM, Essink-Bot ML, Krabbe PF, Bonsel GJ. Test–retest reliability of health state valuations collected with the EuroQoL questionnaire. *Soc Sci Med* 1994;**39**:1537–44.

van Dalen H, Williams A, Gudex C. Lay peoples evaluations of health – are there variations between different subgroups. *J Epidemiol Commun Health* 1994;**48**:248–53.

Williams A. The measurement and valuation of health: a chronicle. Centre for Health Economics Discussion paper 136, University of York, 1995.

Williams, A. The role of the EuroQoL instrument in QALY calculations. Centre for Health Economics discussion paper 130. York: Centre for Health Economics, University of York, 1995.

Studies using the QWB

Study	Patient group	n	Practicality			Reliability	Descriptive validity		Empirical validity
			Timing	Response rate	Completion rate		Content and face	Construct	
Anderson *et al.* (1989)	Five patient groups	1866	–	–	–	Yes[a]	Yes	–	–
Anderson *et al.* (1994)	Terminally ill	204	–	–	–	–	–	–	–
Andresen *et al.* (1995)	Older adults (mean age 72.5 years)	200	17.4 min	68%	93%	–	–	Yes	–
Bombardier *et al.* (1986, 1991)	Rheumatoid arthritis	303	20 min	–	–	–	Yes	Yes	–
Calfas *et al.* (1992)	Osteoarthritis	40	–	–	–	–	–	Yes	–
Fryback *et al.* (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.* (1992)	Chronic renal failure	83	< 10 min	100%?	100%?	–	–	Yes	Yes
Kaplan *et al.* (1976)	Random sample of the general population	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	50	–	–	98%	–	–	Yes	Yes
Manzetti *et al.* (1994)	Lung transplantation	9	–	–	–	–	–	–	–
Mold *et al.* (1992)	Prostate screening	Excluded – did not collect data							
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.* (1987)	Various out- and inpatients	400	18.2 min	–	–	–	Yes	–	–
Tandon *et al.* (1989)	Congestive heart failure	111	–	–	–	–	Yes	Yes	–
Tramarin *et al.* (1992)	AIDS	42	–	–	–	–	–	–	–
Wu *et al.* (1990)	AIDS	31	10 min	–	–	–	–	Yes	–

[a] *Yes indicates the study reports evidence on this criteria (for or against)*

Studies using the Rosser classification of illness

Study	Patient group	n	Method	Practicality			Reliability	Descriptive validity		Empirical validity
				Timing	Response rate	Completion rate		Content and face	Construct	
Bryan *et al.* (1991)	Chiropody	84	HMQ	–	–	–	Yes	Yes	Yes	–
Chan and 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)	End-stage renal failure	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)	Intensive care unit patients	132	HMQ	–	–	–	–	–	–	–
Kind and Gudex (1994)	Random sample of adults	430	HMQ by interview	–	53% agreed to interview	95.5%	–	–	Yes	–
Launois *et al.* (1994)	Low back pain	146	HMQ	–	–	–	–	–	Yes	–
Magee *et al.* (1992)	Abdominal aortic aneurysm	165	HMQ interview	30 min	–	–	–	–	Yes	Yes
Payne and Galland (1995)	Aortic reconstruction	93	HMQ by interview	–	–	–	–	–	–	–
Petrou *et al.* (1992)	Hip and knee replacement	44,159	Self-assessment and observer assessment	–	–	–	Yes	Yes	–	Yes
Rawles *et al.* (1992)	Myocardial infarction	206	Interview	–	–	–	–	–	Yes	Yes
Watkins *et al.* (1990)	Neurosurgery	50	Clinician assessment	–	–	–	–	–	–	–
Whynes and Neilson (1993)	Colorectal cancer	221	HMQ and clinician assessment	–	–	–	Yes	–	Yes	–
Whynes *et al.* (1994)	Colorectal cancer	351	HMQ	–	76–85%	–	–	–	Yes	–
Williams *et al.* (1985)	Coronary artery bypass graft	–	Clinician	–	–	–	–	Yes	–	–

Studies using the HUI

Study	Patient group	n	HUI version	Practicality			Reliability		Descriptive validity		Empirical validity (hypothetical preferences)
				Timing	Response rate	Completion rate	Inter-rater	Retest	Content and face	Con-struct	
Barr *et al.* (1994)	Survivors of therapy for brain tumours	10	II by professionals and parents	–	100%	100%	Yes	–	–	Yes	–
Barr *et al.* (1993)	Survivors of acute lympho-blastic leukaemia	55	II by professionals	1 min	– (100%?)	– (100%?)	Yes	–	–	Yes	–
Billson and Walker (1994)	Survivors of cancer	63	II by pro-fessionals, parents and child	Doctors: 2 min Patients: 5 min	79%	96%	Yes	–	–	–	–
Boyle *et al.* (1994)	General population	555	III by telephone interview	–	91.2%		–	Yes	–	–	–
Boyle *et al.* (1983)	Low birth weight babies		I by home interview	–	–	–	–	–	–	Yes	Yes
Feeny *et al.* (1993a)	Childhood cancer	28	II by professionals	–	100%	100%	Yes	–	–	Yes	–
Feeny *et al.* (1993b)	High-risk acute lymphoblastic leukaemia	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.* (1995)	Low birth weight children	156	II by mapping	–	–	–	–	–	–	Yes	–

Studies using the 15D

Study	Patient group	n	Practicality			Reliability	Descriptive validity		Empirical validity
			Timing	Response rate	Completion rate		Content and face	Construct	
Lonnqvist *et al.* (1994)	Depression	209	–		96% response and completion combined	–	–	–	Yes
Lonnqvist *et al.* (1995)	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 and completed at 2 year follow-up				
Unpublished studies									
Brommel (1990)	Coronary artery bypass graft	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 and 2)	2007	–		96–99%				
In: Sintonen (1995)	Random general population samples	500	–	72%	96–99%	–	–	Yes	–

Studies using the EQ-5D or EQ-6D

Study	Patient group	n	Practicality			Reliability (retest)	Descriptive validity		Valautions (hypothetical preferences)
			Timing	Response rate	Completion rate		Content and face	Construct	
Brazier *et al.* (1996a)	Elderly (> 75 years old)	380	–	99%	> 90%	Yes	–		Yes
Brazier *et al.* (1993)	General population (16–74 years old)	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 ischaemia	180	10 min (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 Angelsenb (1995)	Gastric cancer	26	–	–	–	–	–	–	–
Sculpherc (1993)	Menorrhagia	200	–	–	–	–	–	–	–

[a] All used except Brazier et al. *(1993). [b] Patients assigned to classification and scored by oncologists. [c] The VAS only*

Appendix 4

Supplementary data for chapter 6: references identified by the search of studies comparing non-preference- and preference-based health measures

Anonymous. Association between recombinant human erythropoietin and quality of life and exercise capacity of patients receiving haemodialysis. Canadian Erythropoietin Study Group. *BMJ* 1990;**300**:573–8.

Bakker C, Rutten van Molken M, Hidding A, van Doorslaer E, Bennett K, van der Linden S. Patient utilities in ankylosing spondylitis and the association with other outcome measures. *J Rheumatol* 1994;**21**:1298–304.

Bakker C, Rutten M, van SantenHoeufft M, Bolwijn P, van Doorslaer E, Bennett K, *et al.* Patient utilities in fibromyalgia and the association with other outcome measures. *J Rheumatol* 1995;**22**:1536–43.

Bass EB, Steinberg EP, Pitt HA, Griffiths RI, Lillemoe KD, Saba GP, *et al.* Comparison of the rating scale and the standard gamble in measuring patient preferences for outcomes of gallstone disease. *Med Decis Making* 1994;**14**:307–14.

Borstlap M, Zant JL, van Soesbergen RM, van der Korst JK. Quality of life assessment: a comparison of four questionnaires: for measuring improvements after total hip replacement. *Clin Rheumatol* 1995;**14**:15–20.

Bosch JL, Hunink MGM. The relationship between descriptive and valuational quality-of-life measures in patienst with intermittent claudication. *Med Decis Making* 1996;**16**:217–25.

Bowe TR. Measuring patient preferences: rating scale versus standard gamble. *Med Decis Making* 1995;**15**:283–5.

Busschbach JJ, Horikx PE, van den Bosch JM, Brutel de la Riviere A, de Charro FT. Measuring the quality of life before and after bilateral lung transplantation in patients with cystic fibrosis. *Chest* 1994;**105**:911–17.

Cairns J, Johnston K, McKenzie L. Developing QALYs from condition-specific outcome measures. University of Aberdeen HERU discussion paper 14/91, 1991.

Churchill D, Keown P, Laupacis A, Muirhead N, Sim D, Slaughter D, *et al.* Association between recombinant human erythropoietin and quality of life and exercise capacity of patients receiving haemodialysis. *BMJ* 1990;**300**:573–8.

Churchill DN, Wallace JE, Ludwin D, Beecroft ML, Taylor DW. A comparison of evaluative indices of quality of life and cognitive function in hemodialysis patients. *Control Clin Trials* 1991;**12**:159S–67S.

Fryback DG, Dasbach ED, Klein R, Klein BEK, Martin PA, Dorn N, *et al.* Health assessment by SF-36, Quality of Well-Being Index and time trade-offs: predicting one measure from another. *Med Decis Making* 1992;**12**:348P.

Ganiats TG, Palinkas LA, Kaplan RM. Comparison of Quality of Well-Being Scale and Functional Status Index in patients with atrial fibrillation. *Med Care* 1992;**30**:958–64.

Hornberger JC, Redelmeier DA, Petersen J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *J Clin Epidemiol* 1992;**45**:505–12.

Jaeschke R, Guyatt GH, Willan A, Cook D, Harper S, Morris J, *et al.* Effect of increasing doses of beta agonists on spirometric parameters, exercise capacity, and quality of life in patients with chronic airflow limitation. *Thorax* 1994;**49**:479–84.

Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures. *Q Life Res* 1993;**2**:467–75.

Laupacis A, Wong C, Churchill D. The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. The Canadian Erythropoietin Study Group. *Control Clin Trials* 1991;**12**:168S–79S.

Laupacis A, Muirhead N, Keown P, Wong C. A disease-specific questionnaire for assessing quality of life in patients on hemodialysis. *Nephron* 1992;**60**:302–6 (erratum: appears in *Nephron* 1992;**61**(2):248).

Laupacis A, Bourne R, Rorabeck C, Feeny D, Wong C, Tugwell P, *et al.* The effect of elective total hip replacement on health-related quality of life. *J Bone Joint Surg Am* 1993;**75**:1619–26.

Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;**28**:632–42.

McLeod RS, Taylor BR, O'Connor BI, Greenberg GR, Jeejeebhoy KN, Royall D, *et al.* Quality of life, nutritional status, and gastrointestinal hormone profile following the Whipple procedure. *Am J Surg* 1995;**169**:179–85.

O'Brien B, Viramontes JL. Willingness to pay: a valid and reliable measure of health state preference? *Med Decis Making* 1994;**14**:289–97.

Oldridge N, Guyatt G, Jones N, Crowe J, Singer J, Feeny D, *et al.* Effects on quality of life with comprehensive rehabilitation after acute myocardial infarction. *Am J Cardiol* 1991;**67**:1084–9.

O'Leary JF, Fairclough DL, Jankowski MK, Weeks JC. Comparison of time-tradeoff utilities and rating scale values of cancer patients and their relatives: evidence for a possible plateau relationship. *Med Decis Making* 1995;**15**:132–7.

Read JL, Quinn RJ, Hoefer MA. Measuring overall health: an evaluation of three important approaches. *J Chronic Dis* 1987;**40**(Suppl 1):7S–26S.

Revicki DA. Relationship between health utility and psychometric health status measures. *Med Care* 1992;**30**:MS274–82.

Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Q Life Res* 1993;**2**:477–87.

Revicki DA, Wu AW, Murray MI. Change in clinical status, health status, and health utility outcomes in HIV-infected patients. *Med Care* 1995;**33**:AS173–82.

Rutten Van Molken MPMH, Custers F, Van Doorslaer EKA, Jansen CCM, Heurman L, Maesen FPV, *et al.* Comparison of performance of four instruments in evaluating the effects of salmeterol on asthma quality of life. *Eur Resp J* 1995;**8**:888–98.

Tsevat J, Goldman L, Lamas GA, Pfeffer MA, Chapin CC, Connors KF, *et al.* Functional status versus utilities in survivors of myocardial infarction. *Med Care* 1991;**29**:1153–9.

Tsevat J, Goldman L, Soukup JR, Lamas GA, Connors KF, Chapin CC, *et al.* Stability of time-tradeoff utilities in survivors of myocardial infarction. *Med Decis Making* 1993;**13**:161–5.

Tsevat J, Solzan JG, Kuntz KM, Currier JS, Sell RL, Weinstein MC. Health values of patients with human-immunodeficiency-virus – relationship to mental health and physical functioning. *Med Care* 1996;**34**:44–57.

Visser MC, Fletcher AE, Parr G, Simpson A, Bulpitt CJ. 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. *J Clin Epidemiol* 1994;**47**:157–63.

Zug KA, Littenberg B, Baughman RD, Kneeland T, Nease RF, Sumner W, *et al.* Assessing the preferences of patients with psoriasis. A quantitative, utility approach. *Arch Dermatol* 1995;**131**:561–8.

Appendix 5

Supplementary data for chapter 7

Excluded articles (after review)

MEDLINE database

Author(s)	Reason for exclusion from study
Scheela (1995)	Review article
Lutz and Hallock (1995)	No comparator- not an economic evaluation
Dougherty (1995)	Model, 'illustrative' QALY data only
Wolf *et al.* (1995)	No cost data (resource use only reported) – not an economic evaluation
Hujoel (1995)	No cost data collected – not an economic evaluation
Lok (1995)	No comparator (before and after study) – not an economic evaluation
Munro and Deprest (1995)	No HSM measurement in study
Ennever and Lave (1995)	Review article
Sweet (1995)	No HSM measurement in study
Smith (1995)	Modelling – QWB values and SF-36 transformed from other studies
Merrick (1995)	Review/modelling/methodology – not economic evaluation
Lepoff (1995)	No comparator – not an economic evaluation
McGaughey *et al.* (1995)	No HSM measurement in study
Kind and Sorenson (1995)	No recognised HSM instrument in study
Magotti *et al.* (1995)	QALY values are authors' opinion
Morrison *et al.* (1995)	Review article
Rho and Yoshikawa (1995)	Modelling – QALY values 'assumed'
Chouaid *et al.* (1995)	QALY values are authors' opinion

Other databases

Author(s)	Reason for exclusion from study
Green (1995)	Hypothetical CUA example
Byles *et al.* (1995)	Modelling no QALY data actually collected or derived in study
Duffield *et al.* (1995)	Not an economic evaluation – no comparator
van-Rijswijk (1995)	Modelling – no data collected within study
Whittenburg (1995)	Not an economic evaluation – no costing included
Irani *et al.* (1995)	Modelling – no data collection on either costs or HSM
Sherman and Ryan (1995)	No HSM measurement in study
Karlsson *et al.* (1995)	Cost-effectiveness analysis
Daiker (1995)	Review article
Chrystie *et al.* (1995)	Not an economic evaluation – no comparator
Warren and Rozell (1995)	Review article

Included study references

Cottrell JJ, Openbrier D, Lave JR, Paul C, Garland JL. Home oxygen therapy: a comparison of 2- vs 6-month patient reevaluation. *Chest* 1995;**107**(2):358–61.

Gournay K, Brooking J. The community psychiatric nurse in primary care: an economic analysis. *J Adv Nurs* 1995;**22**:769–78.

Hallstrom AP, Greene HL, Wyse DG, Zipes D, Epstein AE, Domanski MJ, *et al.* Antiarrhythmics versus implantable defibrillators (AVID) – rationale, design, and methods. *Am J Cardiol* 1995;**75**:470–5.

Johnson BF, Evans L, Drury R, Datta D, Morris-Jones W, Beard JD. Surgery for limb threatening ischaemia: a reappraisal of the costs and benefits. *Eur J Vasc Endovasc Surg* 1995;**9**(2):181–8.

Kennedy W, Reinharz D, Tessier G, Constandriopoulos A-P, Traput I, Champagne F. Cost utility of chemotherapy and best supportive care in non-small cell lung cancer. *PharmacoEconomics* 1995;**8**:316–23.

Kerridge RK, Glasziou PP, Hillman KM. The use of 'quality-adjusted life years' (QALYs) to evaluate treatment in intensive care. *Anaesth Intens Care* 1995;**23**:322–31.

Knobbe CA,Carey SP, Rhodes L, Horner RH. Benefit–cost analysis of community residential versus institutional services for adults with severe mental retardation and challenging behaviors. *Am J Ment Retard* 1995;**99**(5):533–41.

Lawrence K, McWhinnie D, Goodwin A, Doll H, Gordon A, Gray A, *et al.* Randomised controlled trial of laparoscopic versus open repair of inguinal hernia: early results. *BMJ* 1995;**311**:981–5.

Mark DB, Hlatky MA, Califf RM, Naylor CD, Lee KL, Armstrong PW, *et al.* Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *New Engl J Med* 1995;**332**:1418–24.

Prince JM, Manley MS, Whiteneck GG. Self-managed versus agency-provided personal assistance care for individuals with high level tetraplegia. *Arch Phys Med Rehab* 1995;**76**:919–23.

The UK Small Aneurysm Trial Participants. The UK Small Aneurysm Trial: design, methods and progress. *Eur J Vasc Endovascu Surg* 1995;**9**(1):42–8.

Uylde Groot CA, Hagenbeek A, Verdonck LF, Lowenberg B, Rutten FFH. Cost-effectiveness of ABMT in comparison with chop chemotherapy in patients with intermediate- and high-grade malignant non-hodgkin's lymphoma (NHL). *Bone Marrow Trans* 1995;**16**:463–70.

Wimo A, Mattson B, Krakau I, Eriksson T, Nelvig A, Karlsson G. Cost–utility analysis of group living in dementia care. *Int J Tech Assess Health Care* 1995;**11**(1):49–65.

Excluded study references

Byles JE, Redman S, Sanson-Fisher RW, Boyle CA. Effectiveness of two direct-mail strategies to encourage women to have cervical (Pap) smears. *Health Prom Int* 1995;**10**:5–16.

Chouaid C, Roux P, Lavard I, Poirot JL, Housset B. Use of the polymerase chain reaction technique on induced-sputum samples for the diagnosis of Pneumocystis carinii pneumonia in HIV-infected patients. A clinical and cost-analysis study. *Am J Clin Pathol* 1995;**104**:72–5.

Chrystie IL, Zander L, Tilzey A, Wolfe CD, Kenney A, Banatvala JE. Is HIV testing in antenatal clinics worthwhile? Can we afford it? *AIDS Care* 1995;**7**:135–42.

Daiker B. Managed care in workers' compensation: analysis of cost drivers and vendor selection. *AAOHN J* 1995;**43**:422–7.

Dougherty CJ. Quality-adjusted life years and the ethical values of health care... reprinted from *Am J Phys Med Rehabil* 1994;**73**:61–5. *Am J Physl Med Rehabi* 1995;**74**(Suppl):S29–33.

Duffield C, Pelletier D, Donoghue J. A profile of the clinical nurse specialist in one Australian state. *Clin Nurse Special* 1995;**9**:149–54.

Ennever FK, Lave LB. Parent preferences and prenatal testing for neural tube defects. *Epidemiol* 1995;**6**:8–16.

Green LG. The comeback kid: pediatric TB. *RT: J Respir Care Pract* 1995;**8**:77–8.

Hujoel PP. Definitive vs. exploratory periodontal trials: a survey of published studies. *J Dent Res* 1995;**74**:1453–8.

Irani J, Fauchery A, Dore B, Bon D, Marroncle M, Aubert J. Systematic removal of catheter 48 hours following transurethral resection and 24 hours following transurethral incision of prostate: a prospective randomized analysis of 213 patients. *J Urol* 1995;**153**:1537–9.

Karlsson G, Teiwik A, Lundstrom A, Ravald N. Costs of periodontal and prosthodontic treatment and evaluation of oral health in patients after treatment of advanced periodontal disease. *Commun Dent Oral Epidemiol* 1995;**23**:159–64.

Kind P, Sorensen J. Modelling the cost-effectiveness of the prophylactic use of SSRIs in the treatment of depression. *Int Clin Psychopharmacol* 1995; **10**(Suppl 1):41–8.

Lepoff RB. Academic medical centers and managed care. *Arch Pathol Lab Med* 1995;**119**:598–9.

Lok AS. Does interferon therapy for chronic hepatitis B reduce the risks of developing cirrhosis and hepatocellular carcinoma? *Hepatology* 1995;**22**:1336–8.

Lutz DA, Hallock GG. Microsurgical transfer of vascularized tissue to close problem wounds. *AORN J* 1995;**62**:234–8, 240, 242–3.

McGaughey MJ, Kiernan WE, McNally LC, Gilmore DS. A peaceful coexistence? State MR/DD agency trends in integrated employment and facility-based services. *Ment Retard* 1995;**33**:170–80.

Magotti RF, Munjinja PG, Lema RS, Ngwalle EK. Cost-effectiveness of managing abortions: manual vacuum aspiration (MVA) compared to evacuation by curettage in Tanzania. *East Afr Med J* 1995;**72**:248–51.

Merrick JC. Critically ill newborns and the law. The American experience. *J Leg Med* 1995;**16**:189–209.

Morrison WB, Schweitzer ME, Wapner KL, Hecht PJ, Gannon FH, Behm WR. Osteomyelitis in feet of diabetics: clinical accuracy, surgical utility, and cost-effectiveness of MR imaging. *Radiology* 1995;**196**:557–64.

Munro MG, Deprest J. Laparoscopic hysterectomy: does it work?: a bicontinental review of the literature and clinical commentary. *Clin Obstet Gynecol* 1995;**38**:401–25.

Rho JP, Yoshikawa TT. The cost of inappropriate use of anti-infective agents in older patients. *Drugs Aging* 1995;**6**:263–7.

Scheela RA. Remodeling as metaphor: sex offenders' perceptions of the treatment process. *Issues Ment Health Nurs* 1995;**16**:493–504.

Sherman DL, Ryan TJ. Coronary angioplasty versus bypass grafting. Cost–benefit considerations. *Med Clin North Am* 1995;**79**:1085–95.

Smith AJ. New guidelines help measure cost-effectiveness. *Minn Med* 1995;**78**:36.

Sweet LE. Vaccine utilization study – Prince Edward Island. *Can J Public Health* 1995;**86**:193–4.

van-Rijswijk L. Frequency of reassessment of pressure ulcers. *Adv Wound Care: J Prevent Heal* 1995;**8**:28/19–28/24.

Warren IB, Rozell BR. Supplemental staffing. Nurse manager views of costs, benefits, and quality of care. *J Nurs Adm* 1995;**25**:51–7.

Whittenburg C. A program proposal for new technology assessment. *AORN J* 1995;**61**:391–2395–9.

Wolf YG, Otis SM, Schwend RB, Bernstein EF. Screening for abdominal aortic aneurysms during lower extremity arterial evaluation in the vascular laboratory. *J Vasc Surg* 1995;**22**:417–21.

Search strategy

No.	Records	Request
1	21,169	COST
2	738	MINIMI?ATION
3	21	COST MINIMI?ATION
4	21,169	COST
5	88,685	EFFECTIVE*
6	389,001	ANALYS*
7	328	COST EFFECTIVE* ANALYS*
8	21,169	COST
9	7127	UTILITY
10	73	COST UTILITY
11	7860	ECONOMIC
12	85,477	EVALUATION
13	191	ECONOMIC EVALUATION
14	562	#3 or #7 or #10 or #13
15	17	'QUALITY-ADJUSTED-LIFE-YEARS'
16	120	QALY*
17	663	#14 or #15 or #16
18	231,661	PY = '1995'
19	130	#17 and (PY = '1995')
20	1,486,389	LA = 'ENGLISH'
21	125	#19 and (LA = 'ENGLISH')
22	4573	COST-BENEFIT-ANALYS*
23	21,169	COST
24	17,842	BENEFIT
25	423,785	ANALY*
26	4666	COST BENEFIT ANALY*
27	4666	#22 or #26
28	1,486,389	LA = 'ENGLISH'
29	4115	#27 and (LA = 'ENGLISH')
30	231,661	PY = '1995'
31	784	#29 and (PY = '1995')
32	814	#21 or #31
33	689	#31 not #21
34	32,475	PT = 'EDITORIAL'
35	56	#32 and (PT = 'EDITORIAL')
36	638	#33 not #35

continued

Search strategy *contd*

No.	Records	Request
37	109,154	PT = 'LETTER'
38	90	#36 and (PT = 'LETTER')
39	548	#36 not #38
40	11,717	PT = 'NEWS'
41	15	#39 and (PT = 'NEWS')
42	533	#39 not #41
43	63,618	PT = 'CLINICAL-TRIAL'
*44	47	#42 and (PT = 'CLINICAL-TRIAL')

Health Technology Assessment panel membership

This report was identified as a priority by the Methodology Panel.

Acute Sector Panel

Current members

Chair:
Professor Francis H Creed, University of Manchester

Professor Clifford Bailey, University of Leeds

Ms Tracy Bury, Chartered Society of Physiotherapy

Professor Collette Clifford, University of Birmingham

Dr Katherine Darton, M.I.N.D.

Mr John Dunning, Papworth Hospital, Cambridge

Mr Jonathan Earnshaw, Gloucester Royal Hospital

Mr Leonard Fenwick, Freeman Group of Hospitals, Newcastle-upon-Tyne

Professor David Field, Leicester Royal Infirmary

Ms Grace Gibbs, West Middlesex University Hospital NHS Trust

Dr Neville Goodman, Southmead Hospital Services Trust, Bristol

Professor Mark P Haggard, MRC

Professor Robert Hawkins, University of Manchester

Dr Duncan Keeley, General Practitioner, Thame

Dr Rajan Madhok, East Riding Health Authority

Dr John Pounsford, Frenchay Hospital, Bristol

Dr Mark Sculpher, University of York

Dr Iqbal Sram, NHS Executive, North West Region

Past members

Professor John Farndon, University of Bristol*

Professor Senga Bond, University of Newcastle-upon-Tyne

Professor Ian Cameron, Southeast Thames Regional Health Authority

Ms Lynne Clemence, Mid-Kent Health Care Trust

Professor Cam Donaldson, University of Aberdeen

Professor Richard Ellis, St James's University Hospital, Leeds

Mr Ian Hammond, Bedford & Shires Health & Care NHS Trust

Professor Adrian Harris, Churchill Hospital, Oxford

Dr Gwyneth Lewis, Department of Health

Mrs Wilma MacPherson, St Thomas's & Guy's Hospitals, London

Dr Chris McCall, General Practitioner, Dorset

Professor Alan McGregor, St Thomas's Hospital, London

Professor Jon Nicholl, University of Sheffield

Professor John Norman, University of Southampton

Professor Michael Sheppard, Queen Elizabeth Hospital, Birmingham

Professor Gordon Stirrat, St Michael's Hospital, Bristol

Dr William Tarnow-Mordi, University of Dundee

Professor Kenneth Taylor, Hammersmith Hospital, London

Diagnostics and Imaging Panel

Current members

Chair:
Professor Mike Smith, University of Leeds

Dr Philip J Ayres, Leeds Teaching Hospitals NHS Trust

Dr Paul Collinson, Mayday University Hospital, Thornton Heath

Dr Barry Cookson, Public Health Laboratory Service, Colindale

Professor David C Cumberland, University of Sheffield

Professor Adrian Dixon, University of Cambridge

Mr Steve Ebdon-Jackson, Department of Health

Mrs Maggie Fitchett, Association of Cytogeneticists, Oxford

Dr Peter Howlett, Portsmouth Hospitals NHS Trust

Professor Alistair McGuire, City University, London

Dr Andrew Moore, Editor, *Bandolier*

Dr Peter Moore, Science Writer, Ashtead

Professor Chris Price, London Hospital Medical School

Dr William Rosenberg, University of Southampton

Dr Gillian Vivian, Royal Cornwall Hospitals Trust

Dr Greg Warner, General Practitioner, Hampshire

Past members

Professor Michael Maisey, Guy's & St Thomas's Hospitals, London*

Professor Andrew Adam, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Dr Pat Cooke, RDRD, Trent Regional Health Authority

Ms Julia Davison, St Bartholomew's Hospital, London

Professor MA Ferguson-Smith, University of Cambridge

Dr Mansel Hacney, University of Manchester

Professor Sean Hilton, St George's Hospital Medical School, London

Mr John Hutton, MEDTAP International Inc., London

Professor Donald Jeffries, St Bartholomew's Hospital, London

Dr Ian Reynolds, Nottingham Health Authority

Professor Colin Roberts, University of Wales College of Medicine

Miss Annette Sergeant, Chase Farm Hospital, Enfield

Professor John Stuart, University of Birmingham

Dr Ala Szczepura, University of Warwick

Mr Stephen Thornton, Cambridge & Huntingdon Health Commission

Dr Jo Walsworth-Bell, South Staffordshire Health Authority

* Previous Chair

continued

continued

Methodology Panel

Current members

Chair:
Professor Martin Buxton, Brunel University

Professor Doug Altman, Institute of Health Sciences, Oxford

Dr David Armstrong, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Professor Nick Black, London School of Hygiene & Tropical Medicine

Professor Ann Bowling, University College London Medical School

Dr Mike Clarke, University of Oxford

Professor Michael Drummond, University of York

Dr Vikki Entwistle, University of Aberdeen

Professor Ewan Ferlie, Imperial College, London

Professor Ray Fitzpatrick, University of Oxford

Professor Jeremy Grimshaw, University of Aberdeen

Dr Stephen Harrison, University of Leeds

Mr John Henderson, Department of Health

Professor Richard Lilford, Regional Director, R&D, West Midlands

Professor Theresa Marteau, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Dr Henry McQuay, University of Oxford

Dr Nick Payne, University of Sheffield

Professor Margaret Pearson, NHS Executive North West

Professor David Sackett, Centre for Evidence Based Medicine, Oxford

Dr PAG Sandercock, University of Edinburgh

Dr David Spiegelhalter, Institute of Public Health, Cambridge

Professor Joy Townsend, University of Hertfordshire

Past members

Professor Anthony Culyer, University of York *

Professor Michael Baum, Royal Marsden Hospital

Dr Rory Collins, University of Oxford

Professor George Davey-Smith, University of Bristol

Professor Stephen Frankel, University of Bristol

Mr Philip Hewitson, Leeds FHSA

Mr Nick Mays, King's Fund, London

Professor Ian Russell, University of York

Dr Maurice Slevin, St Bartholomew's Hospital, London

Professor Charles Warlow, Western General Hospital, Edinburgh

Pharmaceutical Panel

Current members

Chair:
Professor Tom Walley, University of Liverpool

Dr Felicity Gabbay, Transcrip Ltd

Mr Peter Golightly, Leicester Royal Infirmary

Dr Alastair Gray, Health Economics Research Unit, University of Oxford

Professor Rod Griffiths, NHS Executive West Midlands

Mrs Jeanette Howe, Department of Health

Professor Trevor Jones, ABPI, London

Ms Sally Knight, Lister Hospital, Stevenage

Dr Andrew Mortimore, Southampton & SW Hants Health Authority

Mr Nigel Offen, Essex Rivers Healthcare, Colchester

Mrs Marianne Rigge, The College of Health, London

Mr Simon Robbins, Camden & Islington Health Authority, London

Dr Frances Rotblat, Medicines Control Agency

Dr Eamonn Sheridan, St James's University Hospital, Leeds

Mrs Katrina Simister, Liverpool Health Authority

Dr Ross Taylor, University of Aberdeen

Past members

Professor Michael Rawlins, University of Newcastle-upon-Tyne*

Dr Colin Bradley, University of Birmingham

Professor Alasdair Breckenridge, RDRD, Northwest Regional Health Authority

Ms Christine Clark, Hope Hospital, Salford

Mrs Julie Dent, Ealing, Hammersmith & Hounslow Health Authority, London

Mr Barrie Dowdeswell, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Tim Elliott, Department of Health

Dr Desmond Fitzgerald, Mere, Bucklow Hill, Cheshire

Professor Keith Gull, University of Manchester

Dr Keith Jones, Medicines Control Agency

Dr John Posnett, University of York

Dr Tim van Zwanenberg, Northern Regional Health Authority

Dr Kent Woods, RDRD, Trent RO, Sheffield

* Previous Chair

Population Screening Panel

Current members

Chair:
Professor Sir John Grimley Evans,
Radcliffe Infirmary, Oxford

Ms Stella Burnside,
Altnagelvin Hospitals Trust,
Londonderry

Mr John Cairns,
University of Aberdeen

Professor Howard Cuckle,
University of Leeds

Dr Carol Dezateux,
Institute of Child Health,
London

Dr Anne Dixon Brown,
NHS Executive,
Anglia & Oxford

Professor Dian Donnai,
St Mary's Hospital, Manchester

Dr Tom Fahey,
University of Bristol

Mrs Gillian Fletcher,
National Childbirth Trust

Dr JA Muir Gray,
Institute of Health Sciences,
Oxford

Professor Alexander Markham,
St James's University Hospital,
Leeds

Dr Ann McPherson,
General Practitioner, Oxford

Dr Susan Moss,
Institute of Cancer Research

Dr Sarah Stewart-Brown,
University of Oxford

Past members

Dr Sheila Adam,
Department of Health*

Professor George Freeman,
Charing Cross & Westminster
Medical School, London

Dr Mike Gill, Brent & Harrow
Health Authority

Dr Anne Ludbrook,
University of Aberdeen

Professor Theresa Marteau,
Guy's, King's & St Thomas's
School of Medicine
& Dentistry, London

Professor Catherine Peckham,
Institute of Child Health,
London

Dr Connie Smith,
Parkside NHS Trust,
London

Ms Polly Toynbee, Journalist

Professor Nick Wald,
University of London

Professor Ciaran Woodman,
Centre for Cancer
Epidemiology, Manchester

Primary and Community Care Panel

Current members

Chair:
Dr John Tripp,
Royal Devon & Exeter
Healthcare NHS Trust

Mr Kevin Barton,
East London & City
Health Authority

Professor John Bond,
University of Newcastle-
upon-Tyne

Dr John Brazier,
University of Sheffield

Ms Judith Brodie,
Age Concern, London

Mr Shaun Brogan,
Daventry & South Northants
Primary Care Alliance

Mr Joe Corkill,
National Association for
Patient Participation

Dr Nicky Cullum,
University of York

Professor Pam Enderby,
University of Sheffield

Mr Andrew Farmer,
Institute of Health Sciences,
Oxford

Professor Richard Hobbs,
University of Birmingham

Professor Allen Hutchinson,
University of Sheffield

Dr Phillip Leech,
Department of Health

Dr Aidan Macfarlane,
Oxfordshire Health Authority

Professor David Mant,
Institute of Health Sciences,
Oxford

Dr Chris McCall,
General Practitioner, Dorset

Dr Robert Peveler,
University of Southampton

Professor Jennie Popay,
University of Salford

Ms Hilary Scott,
Tower Hamlets Healthcare
NHS Trust, London

Dr Ken Stein,
North & East Devon
Health Authority

Past members

Professor Angela Coulter,
King's Fund, London*

Professor Martin Roland,
University of Manchester*

Dr Simon Allison,
University of Nottingham

Professor Shah Ebrahim,
Royal Free Hospital, London

Ms Cathy Gritzner,
King's Fund, London

Professor Andrew Haines,
RDRD, North Thames
Regional Health Authority

Dr Nicholas Hicks,
Oxfordshire Health Authority

Mr Edward Jones,
Rochdale FHSA

Professor Roger Jones,
Guy's, King's & St Thomas's
School of Medicine
& Dentistry, London

Mr Lionel Joyce,
Chief Executive, Newcastle City
Health NHS Trust

Professor Martin Knapp,
London School of Economics
& Political Science

Professor Karen Luker,
University of Liverpool

Dr Fiona Moss,
Thames Postgraduate Medical
& Dental Education

Professor Dianne Newham,
King's College London

Professor Gillian Parker,
University of Leicester

Dr Mary Renfrew,
University of Oxford

* Previous Chair

continued

National Coordinating Centre for Health Technology Assessment, Advisory Group

Current members

Chair:
Professor John Gabbay,
Wessex Institute for Health Research & Development

Professor Mike Drummond,
Centre for Health Economics, University of York

Ms Lynn Kerridge,
Wessex Institute for Health Research & Development

Dr Ruairidh Milne,
Wessex Institute for Health Research & Development

Ms Kay Pattison,
Research & Development Directorate, NHS Executive

Professor James Raftery,
Health Economics Unit, University of Birmingham

Professor Ian Russell,
Department of Health Sciences & Clinical Evaluation, University of York

Dr Ken Stein,
North & East Devon Health Authority

Professor Andrew Stevens,
Department of Public Health & Epidemiology, University of Birmingham

Past member

Dr Paul Roderick,
Wessex Institute for Health Research & Development

HTA Commissioning Board

Current members

Chair:
Professor Charles Florey,
Department of Epidemiology & Public Health, Ninewells Hospital & Medical School, University of Dundee

Professor Doug Altman,
Director of ICRF/NHS Centre for Statistics in Medicine, Oxford

Professor John Bond,
Professor of Health Services Research, University of Newcastle-upon-Tyne

Mr Peter Bower,
Independent Health Advisor, Newcastle-upon-Tyne

Ms Christine Clark,
Honorary Research Pharmacist, Hope Hospital, Salford

Professor Shah Ebrahim,
Professor of Epidemiology of Ageing, University of Bristol

Professor Martin Eccles,
Professor of Clinical Effectiveness, University of Newcastle-upon-Tyne

Dr Mike Gill,
Director of Public Health & Health Policy, Brent & Harrow Health Authority

Dr Alastair Gray,
Director, Health Economics Research Centre, University of Oxford

Professor Mark Haggard,
MRC Institute of Hearing Research

Dr Jenny Hewison,
Senior Lecturer, Department of Psychology, University of Leeds

Professor Sir Miles Irving
(Programme Director),
Professor of Surgery, University of Manchester, Hope Hospital, Salford

Professor Alison Kitson,
Director, Royal College of Nursing Institute

Dr Donna Lamping,
Senior Lecturer, Department of Public Health, London School of Hygiene & Tropical Medicine

Professor Alan Maynard,
Professor of Economics, University of York

Professor Jon Nicholl,
Director, Medical Care Research Unit, University of Sheffield

Professor Gillian Parker,
Nuffield Professor of Community Care, University of Leicester

Dr Tim Peters,
Reader in Medical Statistics, Department of Social Medicine, University of Bristol

Professor Martin Severs,
Professor in Elderly Health Care, Portsmouth University

Dr Sarah Stewart-Brown,
Director, Institute of Health Sciences, University of Oxford

Professor Ala Szczepura,
Director, Centre for Health Services Studies, University of Warwick

Dr Gillian Vivian,
Consultant, Royal Cornwall Hospitals Trust

Professor Graham Watt,
Department of General Practice, Woodside Health Centre, Glasgow

Professor Kent Woods,
Regional Director of R&D NHS Executive, Trent

Dr Jeremy Wyatt,
Senior Fellow, Health & Public Policy, School of Public Policy, University College, London

Past members

Professor Ian Russell,
Department of Health Sciences & Clinical Evaluation, University of York*

Professor David Cohen,
Professor of Health Economics, University of Glamorgan

Mr Barrie Dowdeswell,
Chief Executive, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Michael Horlington,
Head of Corporate Licensing, Smith & Nephew Group Research Centre

Professor Martin Knapp,
Director, Personal Social Services Research Unit, London School of Economics & Political Science

Professor Theresa Marteau,
Director, Psychology & Genetics Research Group, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Professor Sally McIntyre,
MRC Medical Sociology Unit, Glasgow

Professor David Sackett,
Centre for Evidence Based Medicine, Oxford

Dr David Spiegelhalter,
MRC Biostatistics Unit, Institute of Public Health, Cambridge

Professor David Williams,
Department of Clinical Engineering, University of Liverpool

Dr Mark Williams,
Public Health Physician, Bristol

* Previous Chair

Health Technology Assessment 1999; Vol. 3: No. 9

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.hta.nhsweb.nhs.uk

ISSN 1366-5278