

The Measurement and Valuation of Utility Based Quality of Life

Recommendations from a Review of the Literature

Professor Jeff Richardson

Director, Health Economics Unit,
Centre for Health Program Evaluation

Professor Jan Abel Olsen

Visiting Professor, University of Tromsø, Norway

Dr. Graeme Hawthorne

Deputy Director, Program Evaluation Unit,
Centre for Health Program Evaluation

Duncan Mortimer

Research Fellow, Health Economics Unit,
Centre for Health Program Evaluation

Richard Smith

Senior Lecturer, Health Economics Unit,
Centre for Health Program Evaluation

August, 1999

ISSN 1325 0663

ISBN 1 876662 08 5

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of General Practice and Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE is supported by Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHORS' ACKNOWLEDGMENTS

This paper reports findings from a literature review commissioned by the Commonwealth of Australia Department of Health and Aged Care to promote informed discussion of the measurement and valuation of utility-based quality of life. The review falls within the consultative process of revising the November 1995 edition of *The Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee: including major submissions involving economic analyses*. The views expressed in this paper are not necessarily those of the Department, the Pharmaceutical Benefits Advisory Committee or its sub-committees, and do not necessarily reflect the final outcome of related aspects of the next edition of the PBAC guidelines.

ABSTRACT and SUMMARY

The Commonwealth Department of Health and Family Services is considering options for the update of its guidelines for the preparation of economic analyses which must accompany submissions to the Pharmaceutical Benefits Advisory Committee (PBAC) for the inclusion of new pharmaceuticals on the Pharmaceutical Benefits Schedule (PBS). As part of this process, the Health Economics Unit of Monash University was commissioned to conduct a review of utility-based quality of life measures with a view to determining the feasibility and/or desirability of developing criteria to assess the quality of utility-based quality of life measures. Such criteria were deemed feasible and desirable and the Health Economics Unit specified a set of guidelines for the measurement and valuation of utility weights, based on theoretical argument and empirical findings drawn from the literature.

This paper reports the recommended guidelines arising from the literature review. In-depth discussion of issues surrounding the utility-based quality of life literature, as well as critical reviews of the quality adjusted life year (QALY) and multi-attribute utility (MAU) literatures are given in several companion CHPE working papers. Our overall recommendations are as follows:

- (i) Ideally, 'core' utility measurement should be carried out with an MAU instrument. Holistic measurement should only be used as a supplement.
- (ii) It is recognised that, especially in a short run, research teams may be unable to meet this requirement and, that well-conducted holistic measurement is clearly better than no utility measurement at all.
- (iii) Where possible, holistic measurement should meet the requirements detailed in the guidelines and the persuasiveness of the utility evidence should be, in large part, a function of the adoption of the recommended procedures.
- (iv) The meaning of cost utility ratios calculated with idiosyncratic methods is difficult to interpret and should be discounted.
- (v) The use of an MAU instrument imposes a particular set of characteristics upon the utility measurement. Optional measurement additional to or, in exceptional circumstance, replacing core measurement should therefore adopt the same characteristics. Any optional or supplementary measurement must be carried out in such a way that the information is comparable with core information from other studies.

TABLE OF CONTENTS

Abstract and Summary

1	Introduction	1
2	Aims and Key Issues of the Review	4
	2.1 Aims of the Review	4
	2.2 Key Issues Highlighted by the Review	4
3	'Core' Measurement	7
	3.1 Recommendations for 'Core' Measurement	7
	3.2 Characteristics of 'Core' Measurement	8
4	Optional Measurement	10
	4.1 Acceptable use of the Holistic Approach to Utility Measurement	10
	4.2 Requirements for use of the Holistic Approach to Utility Measurement	10
	4.2.1 Scenario Construction	11
	4.2.2 Scaling Holistic Scenarios	13
5	Additional Recommendations	15
Appendix 1:	Glossary of Terms	16

The Measurement and Valuation of Utility Based Quality of Life

Recommendations from a Review of the Literature

1 Introduction

In January 1993 Australia became the first country to require the economic evaluation of new drugs before their adoption in the Pharmaceutical Benefits Scheme (PBS). The Pharmaceutical Benefits Advisory Committee (PBAC) first issued a set of guidelines for economic analyses in 1990. Since 1990, the guidelines have been periodically revised. The current revision of the guidelines is presently being reviewed by the PBAC. These guidelines are presently being reviewed by the PBAC. As part of this process, the Health Economics Unit of Monash University was commissioned to conduct a review of utility-based quality of life measures with a view to determining the feasibility and/or desirability of developing criteria to assess the quality of utility-based quality of life measures. Such criteria were deemed feasible and desirable and the Health Economics Unit specified a set of guidelines for the measurement and valuation of utility weights, based on theoretical argument and empirical findings drawn from the literature.

This paper reports the recommended guidelines arising from the literature review with only limited discussion. In-depth discussion of issues surrounding the utility-based quality of life literature, as well as critical reviews of the quality adjusted life year (QALY) and multi-attribute utility (MAU) literatures are given in companion working papers. The criteria used in developing guidelines/recommendations based on the review were as follows:

- (i) **That recommendations should result, as far as possible, in measurements that are comparable between different studies.**

The literature reveals such quantitative differences between the results of different scaling techniques, and between the use of different instruments, that standardization and comparability of measurement must be the overriding consideration. Additionally, it is common and desirable for longitudinal pharmaceutical trials to use an MAU instrument to produce dimension specific and/or total health profiles over time. This practice should be strongly encouraged.

(ii) That recommendations should normally reflect the balance of evidence and opinion demonstrated in the literature review.

Where the authors have differed with the balance of opinion this is indicated using the scales below. Recommendations are classified with respect to three characteristics of the literature. The first is the number of authors that have explicitly discussed the issue. This is indicated on the following three-point scale:

-  < 5
-  5 – 10
-  10 +

Second, the three-point scale below characterizes the extent of agreement in the literature:

-  General agreement
-  Some agreement but with at least one significant dissenting opinion
-  General disagreement

Third, recommendations may be consistent with current practice even when this current practice is not explicitly discussed. The following scale indicates the consistency of recommendations with current practice:

-  Indicates consistency with current usage
-  Indicates a difference from current usage
-  Indicates no fixed current approach

(iii) That where there is little agreement the recommendations reflect the assessment of the team. The scales described above jointly describe the extent to which the team has imposed its own assessment;

(iv) That recommendations should take account of the particular needs of the PBAC; and

(v) That the criteria should be coherent, in the sense that different recommendations are not in conflict with each other.

After detailing the aims and scope of the review in Section 2, the recommended guidelines arising from the review are presented in full. Section 3 describes guidelines for 'core' measurement of utility-based quality of life. This 'core' measurement will permit comparison between studies and should be relatively uncontroversial when evaluated against the criteria for developing guidelines outlined above. Section 4 outlines options for measurement, the acceptability or otherwise of which turns on case by case assessment of the evidence and opinion available in the literature within a particular framework of basic social values. Section 5 outlines several additional recommendations arising from the review.

2 Aims and Key Issues of the Review

2.1 Aims of the Review

The general aims of the review and the authors' approach in addressing these aims are briefly outlined below:

- (i) **Whether it is feasible and/or desirable to develop criteria to assess the quality of utility-based quality of life measures.**

There no longer appears to be a serious argument regarding the feasibility of developing such criteria. The entire notion of QALY measurement has been subject to vigorous debate, peaking in intensity in the late 1980's and early 1990's. More recently there has been a de-facto acceptance that QALYs and CUA are, at least in the foreseeable future, an irreversible 'fact of life'. As documented in the review, numerous aspects of the measurement of QALYs are still under discussion. In practice there is, however, sufficient agreement on the *principle* characteristics of measurement to make the adoption of criteria sensible and defensible. It is to be anticipated, however, that these criteria will evolve through time as ethical issues and technical relationships are clarified.

- (ii) **To specify a set of guidelines for the measurement and valuation of utility weights.**

Guidelines for economic evaluation necessarily involve social value judgements. Some of these are widely accepted and almost certainly common between nations. For example, there is universal acceptance of the importance of the patient's view of health status, the use of utility, or strength of preference, to quantify the importance of health status and the importance of duration and change in health status. Beyond this, there are a variety of other factors that may be taken into account in the measurement of QoL. Many of these (for example, the use of age weights or the importance of program-induced anxiety and risk) have the potential to significantly affect the outcome of a CUA, but are not generally included in evaluation studies. The appropriateness or otherwise of incorporating such factors may vary from country to country due to differing social value judgements. For this reason, the recommended guidelines include core measurement about which there should be little debate and which will permit comparison between studies. There are also suggested options for measurement, the acceptability or otherwise of which turns on case by case assessment.

2.2 Key Issues Highlighted by the Review

Following the terms of reference for the review the authors considered each of the issues below and, when appropriate, adopted them as part of the evaluation framework.

- (i) **What each measurement strategy and each scaling method purports to measure.**
- (ii) **Normative and descriptive validity of each measurement strategy and each scaling technique.**

The notion of *normative* validity enters the literature in the debate over the applicability or otherwise of the Von Neumann Morgenstern axioms and the use of the expected utility hypothesis. In sum, when these axioms were shown to be *descriptively* inaccurate some economists reinterpreted them as having *normative* significance. This in turn was used to justify the use of the standard gamble as a gold standard for utility measurement. This view is widely rejected by many economists, including the present authors. ‘Normative validity’, or the normative basis for valuation, is the appropriate subject matter of explicit social judgements and is not an issue which may be determined objectively by ‘economic theory’.

(iii) Generally accepted methods for applying measurement strategies and scaling techniques.

(iv) Empirical comparisons of measurement strategies and scaling techniques (if any) for QALYs, HYE and DALYs.

We have not recommended the general use of HYE and DALYs. Part of the reason for this is that neither have been extensively tested and validated against other measurement techniques. The HYE is, in effect, one of a potentially large number of QALY like measures and the recommendations allow for their use in particular circumstances.

(v) Whether some measures are more/less appropriate in the evaluation of pharmaceuticals.

In principle, the evaluation of a pharmaceutical is no different from the evaluation of any other health-related intervention. In practice, pharmaceutical evaluation commonly involves a longitudinal clinical trial and, in this context, there is a distinct advantage in the use of an MAU instrument. This is explicitly acknowledged and taken into account in the recommendations.

(vi) Review the empirical evidence on converting rating scales into a time trade-off or standard gamble using mathematical transformations.

At present there are no general and accepted transformations and our recommendation is to use such transformations only when they have been validated in the specific context of the particular health states which are the subject of measurement.

(vii) Generally accepted methods for deriving health state descriptions.

The methods suggested in the recommendations reflect current practice and the assessment of the authors. We have not located studies that specifically discriminate between methods.

(viii) Empirical comparisons of mean utility values for different groups of people (patients, partners, health care professionals, men/women, community samples).

Detailed discussion of each of these issues, and critical reviews of the quality adjusted life year (QALY) and multi-attribute utility (MAU) literatures are given in companion CHPE working papers.

3 'Core' Measurement

3.1 Recommendations for 'Core' Measurement

- (i) **An MAU instrument should be employed wherever this is possible.** 📖📖😊❖

The reason for this recommendation is that, as noted above, comparability between studies is the overriding concern. This could be compromised severely by the use of different techniques for the construction and presentation of descriptive systems and through the use of differing scaling techniques. The use of a common MAU instrument allows for in principle verification and replication of the results that have been submitted.

This recommendation implies that studies would not generally employ the holistic approach to QALY construction and should not employ direct measurement.

- (iia) **Studies should employ as first choice (i) the Canadian Health Utility Index (HUI) Mark II or Mark III, the Assessment of Quality of Life (AQoL), or the EuroQoL with British utility weights. Where these are not suitable the 15D should be used.** 📖📖😊❖

- (iib) **Studies should not employ the Rosser Kind Index, the Quality of Wellbeing or the Health Utility Index Mark I.** 📖📖😊✓

Ideally, a single MAU instrument should be employed. In practice, there is no instrument that has been universally accepted or that has been shown superior to all others. The HUI (II), the AQoL and EuroQoL are given first priority as they are based upon acceptable scaling techniques (ie the time trade-off and standard gamble). Use of the AQoL is provisional upon the satisfactory outcome of the current validation exercise. It should be noted that no MAU instrument to date has been subject to a satisfactory validation study (a study which contrasts results with those of a selection of appropriate instruments). Such a study is currently underway at the Health Economics Unit at Monash University. This will include EuroQoL, AQoL and HUI II but not the HUI III. HUI II and AQoL have been examined and adopted by a large number of independent research teams. Results from HUI II but not from AQoL have been published. In sum, there is no gold standard instrument and their status should be periodically reviewed in the light of published information.

Choice between these instruments should be based upon their suitability for (sensitivity to) the measurement needs of particular drugs. We would recommend caution with respect to the use of the EuroQoL. Our review indicates that it is likely to be insensitive to many health states. The onus, however, should be upon authors of submissions to select instruments suitable for their evaluation. The 15D is given lower priority as utility scores are calculated from a rating scale and additive multi attribute model. The Rosser Kind, QWB and HUI (I) are insensitive and produce utility scores suggesting invalidity.

The use of ad hoc utility scales should be unacceptable.

- (iii) **QALY values should be estimated as the discounted value of the future flow of realised (ex post) QALYs.** 📖📖📖☺️✓

This states that for core measurement and comparison, studies should measure the discounted value of current utility as perceived at any point in time without consideration of past or future events. This is equivalent to assuming additive separability of health state utility in different years and in different health states. Thus, in longitudinal trials which track a variable level of current utility during the progression of a disease, the recommended gold standard is the discounted value of the area under the utility-over-time curve. This 'gold standard' clearly embodies a particular set of values that are not universally accepted in the literature (with the first two authors being amongst those who have documented the reasons for including other considerations in utility measurement). The present recommendation is based, firstly, upon current practice and, secondly, upon the need to achieve comparability between core and optional measurement techniques.

- (iv) **Where the treatment effect is to extend survival, studies should report both cost per QALY and cost per life year.** 📖☺️✖️

The trade-off between quality and quantity of life is still somewhat problematical. For this reason, it is important to demonstrate the separate effects of an intervention on quantity and quality.

3.2 Characteristics of 'Core' Measurement

The core recommendations outlined above embody the following measurement characteristics:

- (i) **'Utility' is measured with SG or TTO;** (these being the scaling techniques used by the HUI, AQoL and EuroQoL). 📖📖📖☺️✓
- (ii) **The health state is measured from a patient's perspective,** ie respondents are to imagine that they are in the health state. 📖☺️✓
- (iii) **Utility scores are obtained from a cross section of the general public, not exclusively from patients.** 📖📖📖☺️✓
- (iv) **There is an ex post perspective for hypothetical health states,** ie it is an experienced and not an anticipated health state. 📖☺️✓
- (v) **A single static health state is envisaged not two or more changing health states.** 📖📖📖☺️✓
- (vi) **Prognosis is not taken into account.** 📖📖📖☺️✓
- (vii) **Severity is not taken into account.** 📖📖📖☺️✓

-
- (viii) **‘Potential’ health is not taken into account.** The intention of core measurement is that if the upper limit on a disabled person’s health is less than 1.0, then measured utility gain from an independent illness intervention should not be constrained by this upper limit. We are not aware of any instance where this intention has been contravened. 📖📖😊✓
- (ix) **Age *per sé* (as distinct from life expectancy) is not taken into account.** 📖📖📖😊✓
- (x) **Equity and other concerns are not incorporated in the measurement.** 📖📖📖😊✓

There is no consensus about a number of these characteristics in the literature and, more importantly, their inclusion or exclusion from measurement requires a country specific value judgement. It is desirable that submissions be given some flexibility with respect to the inclusion of these factors and their importance be considered on a case by case basis and for this reason the recommendations below provide guidelines for the exercise of this option.

4 Optional Measurement

4.1 Acceptable use of the Holistic Approach to Utility Measurement

The holistic approach may be employed to supplement—but not normally replace—core measurement in the circumstances listed below.

- (i) **Where the core MAU instruments *demonstrably* do not fit the health state.** Submissions should demonstrate this by comparing results from a sample of respondents using the MAU instruments and an accepted non-utility QoL instrument such as the SF36. 📖📖📖😊❖
- (ii) **Where health states are associated with quantitatively important *ex ante* or process factors** (anxiety, anticipation, risk aversion, fear, hope, dread, etc). 📖😊❖
- (iii) **Where treatment or non treatment is typically associated with a succession of changing health states which have a significant *interactive* effect upon utility and the composite utility is not equal to the sum.** Holistic measurement may capture these synergies and several authors give methodologies for identifying interactive effects. 📖😊❖
- (iv) **Where health outcomes are significantly affected by prognosis.** 📖😊❖
- (v) **Where patient utility scores demonstrably differ significantly from those of the general public.** 📖😊❖
- (vi) **Where it has been impossible for a company to employ core measurement techniques and where results using an holistic approach are available.** 📖😊❖

It is not reasonable to expect the immediate adoption of recommended procedures following distribution of guidelines. However, the use of MAU procedures would normally be far less onerous than the use of the holistic approach and deviation from 'core' measurement should result in a partial discounting of the credibility of the results on the grounds that they are not strictly comparable with other utility measurements.

4.2 Requirements for use of the Holistic Approach to Utility Measurement

- (i) **Whenever the option of supplementary measurement is exercised for one of the reasons discussed above, then the characteristics of 'core' measurement outlined above in Section 3.2 should be followed as far as is possible.**
- (ii) **Where one characteristic has been changed because of its importance for a particular evaluation it should be made clear what the impact of this particular characteristic has been upon the outcome of the evaluation.** 📖📖😊❖

In the exceptional case where it had been demonstrated that variation of two or more of these characteristics would result in quantitatively important differences, then multiple results should be reported which show the impact of changing the characteristics one at a time. For example, it might be argued that an *ex ante* perspective is necessary to capture anxiety and that, because of synergies, multiple health states should be simultaneously evaluated as a health profile. In such circumstances, utility scores should be measured with only one of these characteristics varied (for example, with the simultaneous measurement of multiple states but without anxiety). A second set of results with both characteristics varied should then be measured. The purpose of this recommendation is to permit the separate assessment of the importance of the two atypical characteristics and thereby to allow the PBAC to understand and judge the relevance of the atypical characteristics.

- (iii) **When holistic measurement is included because of an idiosyncrasy of the health state (the exceptional importance of risk prognosis etc) the evidence for this idiosyncrasy should be fully documented.** 📖😊❖

This information is required to demonstrate that unexpected utility results are attributable to the special characteristic of the health state and not to some idiosyncrasy in the utility measurement procedures that have been adopted.

- (iv) **Holistic health states should normally entail a single health state described from an *ex post* perspective.** 📖😊❖

- (v) **Scenarios constructed in application of the holistic approach should not include multiple health states, risk, anticipation, or prognosis.** 📖😊❖

These requirements state that, under normal circumstances, the perspective embodied in the core measurement should be used with optional measurement. When an alternative perspective is (shown to be) of particular significance for a drug then the alternative perspective may be used but this should be accompanied by careful explanation of the need for this perspective and by an assessment of the quantitative effect of the atypical characteristic upon utility scores.

4.2.1 Scenario Construction

- (i) **A complete description of the construction of the scenario employed in the holistic approach should be provided.** 📖😊❖

In order to achieve comparability it is necessary to appreciate the impact of any idiosyncratic features of the methodology. Consequently, deviation from the methods recommended here should be carefully documented.

- (ii) **Holistic scenarios should be constructed on the basis of information provided by patients who have experienced the health state.** 📖📖😊✓

(iii) Scenario construction should not be based entirely upon expert opinion. 📖😊❖

Ideally, the process should involve:

- The collection of information from a group of patients which reflects the range of patient outcomes and experiences;
- The use of focus groups in which representative patients discuss their experience;
- The use of inputs from the focus group to construct initial scenarios;
- Critique of these by relevant professionals (doctor, nurse, etc); and if necessary,
- Feedback of professional information to the patient focus group.

Where focus groups are employed, a skilled facilitator should be employed to ensure that a single person does not dominate the discussion. Multiple focus groups are to be preferred.

(iv) Holistic scenarios should not normally involve more than 5 - 9 attributes or aspects of health states. 📖😊✓

Complex scenarios rapidly lead to cognitive overload. Research indicates that 5 to 9 items of information is the maximum that can normally be understood. For the same reason scenarios should be expressed as simply as possible.

(v) Holistic scenarios may be expressed in the first or third person but should reflect a patient perspective. 📖😊❖

Respondents should be asked to consider that they are the person experiencing the health state.

(vi) Unintended bias due to labelling effects should be avoided. 📖😊❖

Labelling effects—the use of emotive words such as cancer—should be avoided. Effects of the disease should be described entirely in terms of psychological and physical wellbeing, etc.

There is little consensus on this issue and the suggestion could be ignored if it was believed that in a particular circumstance a particular emotive word contributed significantly to the understanding of the health state.

(vii) Submissions should demonstrate an awareness of the potential bias arising from framing. 📖📖😊❖

The economic literature reviewed indicates that these issues are quantitatively significant. It does not present a set of universally acceptable solutions. Researchers should be

encouraged to consult the general psychometric literature and demonstrate that they have taken steps to avoid the most obvious sources of bias from framing and labeling effects.

4.2.2 Scaling Holistic Scenarios

- (i) Utility measurement should be conducted using the time trade-off or standard gamble technique.** 📖📖📖😊✓
- (ii) The rating scale and magnitude estimation techniques should not be used to measure utility.** 📖📖📖😊✓
- (iii) If the person trade-off technique is employed it should be used in conjunction with the time trade-off or standard gamble.** 📖😊✗

There is no gold standard technique for utility measurement and the TTO and SG have not been demonstrated to have a superior 'strong interval' property. However the TTO and SG produce similar results; they are based upon a trade-off involving life and death; they incorporate a personal perspective and are the most commonly employed techniques by economists.

- (iv) Utility scales should be calibrated with zero as death and normal (not perfect) health for a person in the relevant age group as unity.** 📖📖😊✓

Valid scaling requires the use of clear and easily comprehended anchor points for the scale.

- (v) Negative utilities should be included where this is appropriate.** 📖😊✗

A lower boundary of -1.0 should be set for negative utilities. However, where there are a significant number of negative scores sensitivity analysis should be included with this lower bound constrained to -0.5 and -0.2.

- (vi) Utility scores should be obtained from a representative cross section of the general public.** 📖📖📖😊❖

There is disagreement in the literature about whose values should be employed in CUA and there is a strong case for the use of patient values. However, to achieve comparability between generic and holistic measures it is important to use the same values and generic instruments include population values.

- (vii) Utility scores should not be derived by expert judgement.** 📖😊❖

- (viii) Mean not median measures of central tendency should be employed.** 📖😊❖

Utility attempts to measure the strength of preference that requires the use of the mean. The median score indicates the outcome of a vote and does not weight the vote according to the strength of preference. An underlying value judgement here is that each person's capacity for utility is equal.

5 Additional Recommendations

- (i) **When appropriate, studies should employ a condition specific outcome measure. When these cannot be linked to the generic utility instrument they should be linked to changes in utility through the use of holistic techniques.** 📖📖📖😊❖

All of the generic instruments may be insensitive to the differences between different interventions. It is for this reason that the use of a condition specific outcome measure is generally advised, and we endorse this advice. However, for comparability with other studies these measures must be linked to utility and converted into QALYs. There is no standard procedure for doing this. When the generic measure cannot be used we recommend the application of holistic techniques at the baseline of the study to determine utility for CSOM health states. Interpolating between utility levels using the condition specific outcome measure allows approximation of the time profile of utility.

- (ii) **When the patient group consists of, or includes, individuals with a permanent long-term handicap (for example quadriplegia) the utility scale should be recalibrated so that the normal long-term health state has a utility score of unity.** 📖😊✖
- (iii) **QALYs should not receive additional age weights.** 📖📖📖😊✓
- (iv) **Where there is a particularly severe pre-treatment health state the value of this health state may be measured and its influence documented.** 📖😊❖

APPENDIX 1

Glossary of Terms

A1.1 Glossary of Terms

- 15D** A 15 dimensional instrument designed and developed by Harry Sintonen in Finland. This pioneering instrument, developed in the early 1980s remains one of the most sensitive of the instruments currently available. It is scaled using a rating scale and dimensions are combined with an additive model.
- Composite (holistic) utility measurement** This term describes one of the three ways in which utility may be measured. With this approach a complete health state (consisting of multiple dimensions) is described in a vignette or brief narrative describing the health state. This is then used in a survey in which respondents are asked to evaluate and score the health state using one of the utility scaling techniques (TTO, SG, PTO). Utility may also be measured using a multi-attribute utility instrument such as the AQoL or it may be measured directly by asking patients to rate their own health state.
- DALY (Disability adjusted life year)** This is a concept introduced recently by the World Bank/WHO study team lead by Chris Murray and Alan Lopez. The DALY is a QALY like measure in which life years are adjusted by a utility index derived from the PTO. More specifically, PTO values were obtained for 22 health states during a series of focus groups carried out with health experts during which the health states were discussed. The 22 states were located on a rating scale and all other major diseases were then located on the rating scale in relationship to the 22 marker states, again by health experts. Results were used to estimate the burden of disease for all major disease groups in all countries.
- Decision analysis** This is a branch of theory that assists with decision making. In essence, complex decisions are broken down into their constituent parts (a complex health state is broken down into its multiple dimensions). The value or utility of each aspect of the decision (item or dimension of health) is independently estimated (using one of the numerous scaling techniques evolved by decision theorists). The decomposed values or utility scores are recombined using one of several possible models (additive or multiplicative being those that have been adopted by MAU instrument makers).

Descriptive system	This refers to the questionnaire-like set of statements and response categories that define the health domain of a QoL instrument
Dimension	A sub-set of items which define a coherent sub-set of the quality of life. For example, physical health, mental health, social health.
Disease specific instruments	Instruments structured to emphasise dimensions of particular relevance in assessing outcomes for a specific condition. Disease specific instruments are more likely to be sensitive to changes in health status for their target patient group than are generic instruments such as the EQ-5D or HUI.
Economic costs	In principle, economic (or opportunity) costs represent the benefits foregone as a result of any economic activity. The economic cost of reading this glossary is the benefit that may have been obtained from reading something else or undertaking another activity. In practice, economic costs are generally measured by the market price of a commodity where this is available or by an estimated market price equivalent when it is not. Economic costs are compared with dollar benefits in cost benefit analysis and a ranking of the ratio of cost to benefits is obtained in cost effectiveness analysis and cost utility analysis.
EuroQoL or EQ-5D	This MAU instrument consists of six items that define 243 response categories (Dolan, 1997). The instrument was designed for cross European comparisons of the quality of life. It was subsequently the basis for one of the largest scaling exercises to date, using the time trade-off technique. The research, carried out by Alan Williams and his colleagues at York, resulted in several sets of 'utility tariffs' each applying to a different age, sex and SES group.
Factor analysis	Factor analysis collapses many items or variables into a few underlying dimensions or constructs. That is, several items assessing hearing, speech, and vision might load highly on a factor that captures the underlying HRQoL dimension of 'Physical Senses'. The association between items and factors determines the semantic content of dimensions. Factor analysis allows the structure of an instrument to be verified or empirically derived.
Health Utility Index (HUI)	The HUI Mark I, II and III are three MAU instruments developed by George Torrance and his colleagues at McMaster University. Along with the AQoL, the HUI instruments are the only ones to employ a multiplicative model for combining utilities.

HRQoL (Health related quality of life)	This is a relatively loose concept with refers to the sub-set of quality of life which may be affected by a health program. As health programs may, potentially, affect all aspects of life, HRQoL could, arguably, be considered identical with the concept of quality of life.
Healthy year equivalent (HYE)	This is one particular form of QALY which is derived by using the standard gamble to evaluate, not one but all of the years of life that will be experienced in a given health state. The concept has been vigorously promoted by Gafni at McMaster University. There has been a major debate to determine whether or not the HYE is conceptually different from measurement over multiple years using the TTO.
Instrument	This term has its normal meaning <i>vis</i> a device for achieving a particular objective. In the present context the device is the descriptive system all questionnaire – type set of items with which a respondent can indicate – describe – their health state. The instrument may or may not have a corresponding set of utility weights. If it does not, it is commonly described as a 'psychometric instrument' (ie where unweighted response scores can be summed). Where multiple dimensions are described and utility weights or scores are included, it becomes 'a multi-attribute utility' instrument.
Item	A single statement dealing with a single element or concept. Each item typically has multiple response categories (eg: not at all, a little, a lot, a great deal).
Latent variable	An unobserved variable which is postulated as a construct and is defined by a sub-set of items. Each health dimension represents a latent variable as does the global concept of quality of life.
Orthogonality (Independence)	Decision analytic theory, upon which the modelling (combination) of utility scores is based requires that the different dimensions of health should be orthogonal (uncorrelated) with each other. This independence is also known as structural independence. (See structural independence).
Person Trade-off (PTO)	This scaling device was once known as the equivalence technique but has been (more descriptively) redefined and popularised by Erik Nord. The technique asks respondents to consider two programs. The first returns a defined number (eg 100) patients from imminent death to the health state being evaluated). The second program returns a variable number, N, patients from imminent death to full health. N is varied until the

two programs appear equally attractive. It is then possible to infer that the utility of the health state is equal to $n/100$.

Preference independence This refers to the requirement of multi-attribute utility theory that the utility value of an item should not depend upon a person's position on an unrelated item or combination of items. Preference independence would be violated if the dis-utility of anxiety were greater when anxiety occurred in combination with pain. There are three forms of preference independence (see Keeney and Raiffa, 1976; Winterfeldt and Edwards, 1993; or Feeney et al 1995).

Psychometrics This is the branch of psychology that deals with the quantification of psychological quantities such as intelligence. It is derived from psychophysics that deals with the quantification of the strength of perception of physical stimuli such as heat, light and sound. The psychometric entities dealt with by psychometrics are considered to be a psychological 'construct' and, as such, are analogous to the construct we call 'quality of life'. As these constructs cannot be directly observed, various techniques have evolved for their measurement and validation. All of the original MAU instruments employed psychometric techniques.

Quality adjusted life year (QALY) This is almost a self-explanatory term. QALYs are a metric obtained by multiplying the number of calendar years of life by an index number that reflects the utility of strength of preference for the health state of the person involved. Strictly speaking, the metric should be described as a 'utility adjusted life year' or as 'healthy year equivalents'. The term may either be used narrowly to refer to life years multiplied by an index number which refers to average utility or the term may be used more broadly to refer to the generic set of metrics which involve the adjustment of life years. The broader concept subsumes: 'healthy years equivalents', DALYs, years adjusted by any of the utility measuring devices, utility measured before or after the health state has been experienced, utility of a defined number of years measured as a single outcome, or the utility of a health state that varies through time.

Quality of Well Being (QWB instrument) Along with the Rosser-Kind index this was the earliest of the MAU instruments and was based upon an ad hoc survey of quality of life. It is based upon a rating scale and an additive model. The QWB was the basis for the utility scores used in the State of Oregon to rank all of the procedures available for the poor through the State's Medicaid program.

Rating scale

This is the scaling technique that has been most commonly used by psychometricians. Typically it presents the respondents with a vertical or horizontal line which is calibrated between zero and 1 or zero and 100 where the end points are clearly and unambiguously defined (usually as normal health and death). Respondents are asked to indicate how far between the end points the health state being evaluated should be. Respondents are typically urged to consider the distance on the rating scale to represent the strength of their feeling. Rating scales are seldom used by economists because of the ambiguity in the interpretation of the results.

Standard gamble (SG)

This is a scaling technique that is considered by some to be the gold standard because it employs the axioms of von Neuman and Morgenstern. The standard gamble consists of a choice between (typically, a life time) in the health state of interest and a gamble between life and death in which the probability of life is varied until the gamble is equally attractive as the certainty of life in the inferior health state. At this point of equivalence, the probability of the favourable outcome is taken as an index of the strength of preference (or utility).

Scaling

The process whereby a set of numbers representing relative importance is attached to a descriptive system. If an appropriate scaling technique is adopted (time trade-off, standard gamble) then the scale values may claim to represent utility weights.

Structural independence

This is a requirement of decision theory that is the basis for combining or modelling the utility scores of different health dimensions. Structural independence – the lack of correlation between dimensions – is required so that a particular aspect of health will not be, in effect, double counted. For example, if strong pain was always associated with anxiety (these items were correlated in a patient survey) then the importance of anxiety could be measured both directly and indirectly through the disutility assigned to pain.

Time Trade-off (TTO)

A scaling technique to derive utility weights. In the typical time trade-off, the respondent is offered a choice between the health state of interest for a defined number of years (typically the respondent's life expectancy) and a lesser period of full or normal health. The second option is varied until the respondent indicates that the two options are of equal value. At this point, the utility of the health state may be defined as the ratio of years of full health divided by the larger number of years in the health state.

Utility

An index of the strength of a person's preference. Initially, and as envisaged by Jeremy Bentham, John Stuart Mill and the Victorian Utilitarians, utility referred to the psychological dimension of pleasure/pain. From the 1930s onwards this concept was largely replaced in economics by the concept of preference utility. In philosophy, the original concept similarly competes with preference based utility, although this concept is somewhat different from the economist's concept.

WHO QoL

A large generic instrument designed by a cross-national team coordinated by the WHO. The full WHOQoL is much broader than a conventional MAU instrument. The WHOQoL (brief) is available for use as a psychometric (unweighted) instrument. Utility weights have not yet been derived.