

**Using the Assessment of Quality of Life
(AQoL)**

Version 1

Dr Graeme Hawthorne

Deputy Director and Senior Lecturer,
Program Evaluation Unit, Centre for Health Program Evaluation

Professor Jeff Richardson

Professor and Director,
Health Economics Unit, Centre for Health Program Evaluation

Neil Day

Principal Research Fellow,
Centre for Health Program Evaluation

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
Web Address <http://ariel.unimelb.edu.au/chpe/>

ACKNOWLEDGMENTS

This project was made possible by a project grant from the Victorian Health Promotion Foundation.

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.

AUTHOR ACKNOWLEDGMENTS

We would like to acknowledge all those who contributed to the AQoL project. In particular we would like to thank Dr Richard Osborne for his excellent work during the early phases of the AQoL development, and Ms Helen McNeil who set up the questionnaires and supervised our interviewers.

We would like to acknowledge the generous cooperation and support of St Vincent's Hospital and the Austin and Repatriation Medical Centre. This thanks is extended to the ward and outpatient staff, including the nurses who were so helpful to our interviewers, often at times which were inconvenient for hospital routines.

Our thanks are especially extended to our interviewers who did such a terrific job of contacting and visiting the respondents: Sonia Barthelmebs, Kerith Culley, Janet Day, Marina Hawthorne, Ivan Marsetti, Jan Memery, Noelle Perry and Di Press. Their high level of professionalism and their willingness to work at times to suit the respondents all contributed to the success of this study.

We would also like to thank the respondents for completing interviews and questionnaires which were very long, and which were challenging because of the questions which probed those parts of their lives which were, for many respondents, associated with the losses which accompany illness.

This research was funded by the National Health and Medical Research Council, the Victorian Health Promotion Foundation, the Australian Research Council, and the Vice-Chancellor's Fund at Monash University. Ethics approval was given by the Ethics Committees at The University of Melbourne, Monash University, St Vincent's Hospital, and the Austin & Repatriation Medical Centre. Our thanks are extended to all these bodies for their support.

Finally, we would like to thank Dr Stuart Peacock for his valuable reading of the manuscript.

Dr Hawthorne's position at the University of Melbourne is funded by the Victorian Consortium for Public Health.

Table of Contents

Introduction	i
Glossary of Terms	ii
1 The AQoL Instrument	9
1.1 AQoL construction.....	11
1.2 AQoL validation.....	13
1.3 Utilities.....	17
2 Instrument administration	21
2.1 Mode of administration.....	21
2.2 The time frame	22
2.3 Items which may cause difficulty	22
2.4 Missing data	24
2.5 Using proxies	29
2.6 Sample size.....	29
2.7 Languages other than English	34
2.8 Changing AQoL items.....	34
3 Scoring the AQoL	35
3.1 Data entry and initial coding.....	35
3.2 As an unweighted, psychometric instrument providing 'value' profiles	36
3.3 As a utility HRQoL measure.....	37
3.3.1 Calculating utilities.....	38
3.3.2 Using the AQoL to estimate QALYs.....	41
4 AQoL: Questions and Answers	41
4.1 AQoL and economic evaluation.....	41
4.2 Rationale for the AQoL	44
4.3 The AQoL and other QoL instruments.....	46
4.4 Scope and coverage of the AQoL.....	48
4.5 Validation and psychometric properties.....	49
4.6 QoL and health promotion	51
4.7 Ethical issues	52
References	56
Appendix 1 The AQoL instrument	60
Notes.....	60
The AQoL instrument: Version A.....	61

The AQoL instrument: Version B.....	65
Appendix II AQoL dimension disutility look-up tables.....	71
Appendix III AQoL utility SPSS data analysis program	77

List of Figures

Figure 1	HRQoL dimensions.....	3
Figure 2	AQoL factorial structure.....	4
Figure 3	HRQoL coverage.....	5
Figure 4	Structural equation analysis.....	6
Figure 5	AQoL/Barthel index.....	7
Figure 6	AQoL/Affect balance scale.....	7
Figure 7	AQoL/SF-36.....	8
Figure 8	Item dis-utility values.....	9
Figure 9	AQoL utility formulae.....	10
Figure 10	Mail vs. telephone administration.....	12
Figure 11	Sample size.....	24
Figure 12	Typical AQoL scores by hospital status.....	27
Figure 13	Look-up table: worked example.....	27
Figure 14	Using look-up tables: independent living.....	28
Figure 15	Sample AQoL utility data analysis.....	30

Introduction

This manual is to assist researchers using the *Assessment of Quality of Life* (AQoL) instrument. The AQoL is a multi-attribute utility health-related quality of life instrument for use as a utility instrument for the economic evaluation of programs; it can also be used as an unweighted psychometric tool in program evaluation.

Since its release, use of the AQoL has rapidly grown; at the time of writing 51 research teams had registered with us. Although we support researchers using the AQoL, there is now considerable demand for our services to answer a myriad of questions ranging from issues surrounding instrument selection to how to score the AQoL.

This manual was written with the needs of these researchers in mind. Much of the advice in it is based on our experience of questions asked by researchers using the instrument. The manual therefore attempts to respond to these common needs.

As use of the AQoL increases, researchers will undoubtedly raise other questions and develop solutions to them. In order to improve the advice to researchers, we would be most grateful if users who identify strengths or limitations (and ways of solving these!) could keep us informed so that we can provide the best possible support for all users.

It is our hope that the AQoL meets the need for simple, straightforward, sensitive, valid and reliable health-related quality of life measurement.

Further information on the AQoL itself, issues in research design, administration, data analysis or interpretation can be readily obtained by contacting the AQoL team at the following address:

Mail: Dr Graeme Hawthorne
Centre for Health Program Evaluation
PO Box 477
West Heidelberg, 3044

Telephone: +61 3 9496 4434
Facsimile: +61 3 9496 4424
Email: chpe@buseco.monash.edu.au
g.hawthorne@gpph.unimelb.edu.au

Internet: This manual, copies of the AQoL, the scoring algorithm and our working papers relevant to the AQoL can be found at the CHPE website:
<<http://ariel.unimelb.edu.au/chpe/>>

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 the dimensions are combined with an additive model.

Composite (holistic) utility measurement

This phrase 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. 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 by the World Bank/WHO study team lead by Murray & Lopez. The DALY is a QALY-like measure in which life years are adjusted by a utility index derived from the PTO. More specifically, the World Bank study elicited PTO values 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. Placement was again carried out 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 which 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 then independently estimated (using one of the numerous scaling techniques evolved by decision theorists). The decomposed value 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 in a utility instrument which define the health states of the 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. Different instrument developers have used different names to describe dimensions, including domains, facets and elements.

Disease specific instrument

Instrument for measuring the outcomes from interventions, where the instrument only covers those aspects of HRQoL affected by the disease of interest.

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.

EQ5D/EuroQol

This MAU instrument consists of five items, each of which has three health levels. The instrument was designed for cross-European comparisons of 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 University, resulted in several sets of 'utility tariffs' each applying to a different age, sex and SES group.

Factor analysis

Procedure for examining the underlying structure of responses to items through examination of the correlations. Widely used in instrument

construction to identify which items form logical and statistically consistent scales.

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 which refers to the sub-set of quality of life which may be affected by a person's health status. As health programs may, potentially, affect all aspects of life, HRQoL could, arguably, be considered identical with the concept of quality of life. The distinction, however, is usually taken to imply that HRQoL does not measure the environment in which a person lives or their socio-economic status.

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 the measurement of multiple years using the TTO.

Instrument

This term has its normal meaning; *viz.* a device for measuring a particular construct or object. In the present context the device is the descriptive system of a questionnaire. The items enable a respondent to describe their health status. 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' (i.e. 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 question dealing with a single concept. Comprises two parts; the 'item stem' which defines the 'item content', and the 'item responses' to which the respondent responds. Closed items typically have multiple response categories (e.g. not at all; a

little, a lot, a great deal). Open items are not used in HRQoL measurement.

Latent variable

An unobserved variable which is postulated as a construct and is defined by a set of items. Each health dimension represents a latent variable, as does the global concept of quality of life.

Multi-attribute utility (MAU) instrument

A specialized form of instrument consisting of a set of items and a corresponding set of scores which can be combined into a single index. MAUs ask people to indicate for each item or health-related statement which response most closely corresponds with their own health status. An instrument has multi-attributes if it is a generic instrument which can describe many dimensions of health. It is also a utility instrument if the importance weights attached to each response have been derived using a technique for utility measurement (the standard gamble, the time trade-off, the person trade-off, magnitude estimation).

Orthogonality (independence)

Psychometric requirement for unidimensional measurement, where the measurement of interest can be shown to be statistically uncorrelated with other measurements. This is investigated through factor analysis. 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 'equivalents 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 (e.g. 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. At this point it is possible to infer that the utility 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 was greater when anxiety occurred in combination with pain. There are three forms of preference independence. For details see Keeney & Raiffa (1976) or Winterfeldt & Edwards (1986).

Psychometrics

This is the branch of psychology which deals with measurement. The entities dealt with are considered to be 'constructs' and, as such, are analogous to the construct 'quality of life'. As these constructs cannot be directly observed, various techniques have evolved for their measurement and validation.

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 which reflects the utility or 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 year 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 which varies through time.

Quality of well being (QWB)

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 employs 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. Recently a new version of the QWB has been released.

**Rating scale
(Visual analogue scale: VAS)**

This is a scaling technique commonly used to assess an individual's health. Typically it presents respondents with a vertical or horizontal line which is calibrated between 0–100 where the end points are clearly and unambiguously defined (usually as 'full health' and 'death'). Respondents are asked to indicate how far between the end points they feel the health state being evaluated should be. Respondents are 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 which 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 lifetime in the health state of interest and a gamble between good health (full life) and death, in which the probability of full 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 (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 which is the basis for combining or modelling the utility scores of different health dimensions. Structural independence – the lack of statistical 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 then the importance of anxiety could be measured both directly and indirectly through the dis-utility assigned to pain.

Time trade-off (TTO)

A scaling technique to derive utility weights. In the typical TTO the respondent is offered a choice between the health state of interest for a defined number of years (e.g. the respondent's life

expectancy) and a lesser period of full or good 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 other 19th Century 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.

WHOQoL

A generic instrument with 100-items, designed by a cross-national team coordinated by the World Health Organization. The full WHOQoL is much broader than a conventional MAU instrument. The WHOQoL-Bréf (25-items) is available for use as a psychometric (unweighted) instrument. Utility weights have not yet been derived.

1 The AQoL Instrument

The Australian Quality of Life Project was undertaken to construct and validate a health-related quality of life instrument which would:

1. be a psychometrically appropriate instrument for the evaluation of a range of health interventions, from the medical and pharmacological treatment of acute illness through to health promotion activities; and to
2. enable the economic evaluation of programs through the computation of utilities¹ before and after health-related interventions.

The Project resulted in the construction of the *Assessment of Quality of Life* (AQoL) instrument, a utility instrument comprising 5-scales, and 15-items (Richardson 1994; Hawthorne & Richardson 1995; Hawthorne, Richardson & Osborne 1999).

The recent and growing interest in valid and sensitive health-related quality of life (HRQoL) measurement can be traced to three key interacting factors: changes in the predominant types of diseases from infectious to chronic; the introduction of sophisticated medical technology capable of both saving lives and prolonging the lives of the ill; and growing awareness of the limited resources that are available in the health sector (Nordenfelt 1994). The first two factors have increased the importance of HRQoL as an output. The last factor has meant that we must evaluate and discriminate between the increasingly large number of procedures that are available. This implies the need for output measures that are sensitive to HRQoL and that are suitable for use in both summative and economic evaluation.

The *Assessment of Quality of Life* (AQoL) instrument was designed to assist with meeting these two evaluation requirements. The aims were for an instrument which would:

- a) cover the full universe of HRQoL as far as was practicable;
- b) meet standard psychometric requirements for reliable and valid measurement;
- c) be sensitive to a wide range of health states;
- d) be available as a psychometric instrument (yielding 'health state' scores); and
- e) be capable of use as an economic instrument (yielding 'preference' scores); and

¹ A distinction is sometimes drawn between 'utilities', derived from the standard gamble technique and 'preferences' derived from the time-trade off (TTO) or other techniques which do not involve measurement 'under risk'. The validity of this distinction has been challenged (eg Winterfeldt and Edwards (1986) and Richardson (1994)). In this paper the term 'utility' is reserved for use with the TTO data, and 'value' is used to describe the non-weighted psychometric data. The AQoL can provide each type of measurement (utility and value). For values, these can be described at the dimension level (thus providing a health status profile based on the five dimensions of the AQoL) or at the AQoL HRQoL level (a single value based on summation of dimension scores). For utilities, the AQoL provides a preference profile based on the four of the five dimensions' utilities (see Figure 10) and an overall AQoL utility score (Figure 11).

f) be simple to apply.

In keeping with these aims, the AQoL was developed using standard psychometric procedures as well as being weighted for use in cost utility analysis (CUA). To overcome the limitations associated with valuing composite health states, the AQoL was constructed as a decomposed multi-attribute instrument.

This section provides a brief overview of the construction and validation of the AQoL so that potential users are aware of its underlying properties.

Some issues in psychometric validity

The validity of values obtained from HRQoL-utility instruments depends upon the validity of the descriptive system used and the validity of the numeral values corresponding with each health state that is described. As the accuracy of the numerical values cannot normally be established by direct observation or by comparison with some gold standard, instrument reliability and validity should be determined by establishing a 'nomological net' of evidence assessing the degree of confidence and consequent inferences a researcher can place on the scores obtained by study participants (Cronbach & Meehl 1955; Streiner & Norman 1995).

This net includes instrument content (does it adequately reflect the theoretical universe upon which it is premised?), criteria (how well does the measurement match values obtained from other instruments purporting to measure the same thing?) and construct validity (does the instrument provide scores from which inferences about the defined universe can be made?). Validation of such instruments is an ongoing process and an instrument may prove to be valid in one context but not in another.

Utility weights for economic evaluation

Before the development of cost utility analysis, economic evaluation of health services typically ignored HRQoL or treated it as an 'intangible' that could be noted and described but not quantified or included as an integral part of the health outcome in evaluation research. Cost utility analysis (CUA) attempts to overcome this by adopting the quality adjusted life year (QALY) as the unit of output for health benefits (Torrance 1986).

QALY values have been obtained using one of two broad approaches. With the composite approach, health state scenarios—or complex health state descriptions—have been constructed and numerical values placed upon them using a direct scaling technique. Alternatively, a decomposed approach has been employed using a multi-attribute utility (MAU) instrument. This consists of a generic 'descriptive system' or 'descriptive instrument' and a set of scale values corresponding with each possible health state in the descriptive instrument. As with the scenario-based approach, the values are obtained using one of the standard scaling techniques, viz. magnitude estimation, time trade-off, standard gamble, or, most recently, the person trade-off (Murray & Lopez 1996).

The steps used to construct the AQoL involved:

-
- i) Generating a hypothetical HRQoL model & the various associated dimensions;
 - ii) Developing an item pool to match the dimensions;
 - iii) Administering the item pool to a construction sample;
 - iv) Analyzing the construction sample data to generate an empirical model;
 - v) Comparing the hypothesized & empirical models;
 - vi) Selecting the final model;
 - vii) Scaling it using time-trade off procedures to obtain utility values; and
 - viii) Validating the final instrument.

Section 1.1 briefly reports steps i–vi; some preliminary validation evidence (step viii) is reported in Section 1.2, while Section 1.3 reports on step vii. For more detail on the construction and weighting of the AQoL readers should familiarise themselves with Hawthorne et al (2000) Construction and Utility Scaling of the Assessment of Quality of Life (AQoL) Instrument. For a comparison of the AQoL with other utility instruments readers are referred to Hawthorne et al (2000) Life and Death: Theoretical and Practical Issues in Utility Measurement, CHPE Working Paper 102, Hawthorne et al 2000; Hawthorne et al 2000.

Copies of these papers can be found at the CHPE website <http://ariel.unimelb.edu.au/chpe/>.

1.1 AQoL construction

The project commenced with a literature review of the key HRQoL instruments published since the early 1970s. Copies were obtained and subjected to critical analysis. The results suggested twenty aspects of life were important in measuring HRQoL (see Figure 1). A model was subsequently constructed comprising the HRQoL universe, and the five primary dimensions contributing to this universe (illness, independent living, physical ability, psychological wellbeing and social relationships). A pool of items was generated from the literature, interviews and focus groups with 24 clinicians from St Vincent's Hospital (Melbourne) and the Department of Public Health and Community Medicine at The University of Melbourne. Typical items can be found in the AQoL instrument in Appendix I.

Figure 1 **HRQoL dimensions**

Relative to the body	Social expression
- Anxiety/Depression	- Activities of daily living
- Bodily care	- Communication
- Cognitive ability	- Emotional fulfilment
- General health	- Family role
- Memory	- Intimacy/Isolation
- Mobility	- Medical aids use
- Pain	- Medical treatment
- Physical ability/Vitality	- Sexual relationships
- Rest/Fatigue	- Social function
- Sensory functions	- Work function

Following editing and revision of items, the item pool was administered to a construction sample comprising two cohorts: a list sample of 143 patients from St Vincent's Hospital and a random sample of 112 Melbourne residents selected from the telephone directory.

Standard psychometric procedures were used to examine item properties, and items failing to meet specified criteria were discarded. The remaining items were then pooled and a two-stage factor analysis (principal components and varimax) was used to identify redundant items. Reliability analysis was also carried out. These steps were repeated until the most parsimonious solution was derived consistent with psychometric and measurement theory (Anastasi 1976; Rummel 1970; Pedhazur & Schmelkin 1991).

This resulted in an instrument where all items loaded >0.30 on a principal components analysis, suggesting the instrument measures a single underlying construct. In addition five factors were identified, each with three items, as shown in Figure 2. In this figure the columns are the factors and the rows the individual items. For clarity, each resulting scale has been labelled. The average factor item loadings were 0.74 and on cross-factors they were 0.13; these data indicate the five factors were orthogonal to each other, and that each comprised a single scale. That all items loaded on a principal component leads to the conclusion that scores on the five factors can also be combined, thus providing an overall index. The internal consistency of the instrument was appropriate (Cronbach $\alpha = 0.80$).

Figure 2 AQoL factorial structure

Illness	0.86	0.15	0.09	0.07	0.18
	0.86	0.14	0.08	0.07	0.08
	0.84	0.24	0.07	0.06	0.08
Independent living	0.08	0.87	0.07	0.17	-0.02
	0.15	0.71	0.11	-0.03	0.17
	0.27	0.76	0.08	0.14	0.13
Social relationships	0.02	0.06	0.84	0.12	-0.11
	0.10	0.08	0.71	0.08	0.19
	0.16	0.47	0.56	-0.06	0.18
Physical senses	0.14	0.09	-0.05	0.67	-0.15
	0.14	0.08	0.14	0.79	0.13
	-0.13	0.03	0.14	0.68	0.35
Psychological well-being	0.08	0.07	0.11	0.07	0.75
	0.11	0.01	0.46	0.16	0.64
	0.20	0.24	-0.16	0.04	0.53

1.2 AQoL validation

The literature suggests three inter-related forms of validation—content, construct and criterion—are accepted as providing evidence of the nomological net necessary for accepting that a measure possesses validity (Cronbach & Meehl 1955; Pedhazur & Schmelkin 1991; Anastasi 1986).

Content validity refers to the relationship between the hypothesized universe and the measurement: the measurement must provide adequate coverage of the universe. Following the procedures outlined by Lennon (Lennon 1965), the content of each AQoL item was mapped against the HRQoL universe defined through the literature review. The results are given in Figure 3, along with those of several other popular HRQoL utility instruments (the EuroQoL/EQ5D (EuroQoL Group, 1990), HUI-III (Torrance et al 1995; Torrance 1996; Feeny, Furlong & Torrance 1996) and 15D (Sintonen 1994; Sintonen 1995) and a standard health profile instrument (the SF-36 (Ware et al 1993)). This comparison shows that the AQoL provides good coverage across the important HRQoL dimensions; coverage which is at least as good, if not better than, comparable instruments.

Figure 3 HRQoL coverage

HRQoL dimensions	SF-36	AQoL	EuroQoL	HUI-III	15D
Relative to the body					
Anxiety/Depression	***	*	*		**
Bodily care	*	*	*	*	
Cognitive ability				*	*
General health	*****				
Memory				*	
Mobility	***	*	*	*	*
Pain	**	*	*	**	*
Physical ability/Vitality	*****			*	*
Rest and fatigue	**	*			*
Sensory functions		**		****	*****
Social expression					
Activities of daily living		*	*		*
Communication		*		**	*
Emotional fulfilment	**			**	
Family role		*			
Intimacy/Isolation		*			
Medical aids use		*			
Medical treatment		**			
Sexual relationships					*
Social function	**	*			
Work function	**				

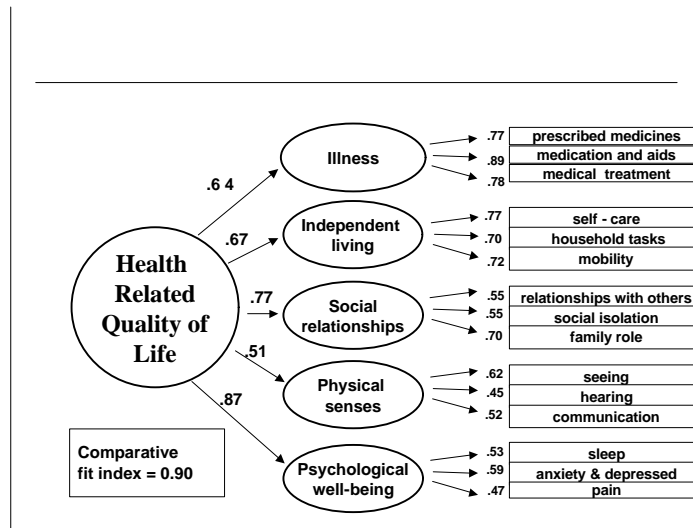
Construct validity refers to how well an instrument's score can be used to infer scores about the underlying psychometric universe or concept that is to be measured. Generally, construct validity is established by examining how well empirical data 'fits' the hypothesized model.

In order to understand the AQoL model it was subjected to structural equation modelling (SEM) (Pedhazur & Schmelkin 1991; McArdle 1996) using a *total disaggregation second order SEM model*, in which each group of items was used to operationalize its respective hypothesized latent dimension. Under these stringent requirements the measures of 'fit'—i.e. estimates of how well a specified model fits the data—typically provide values (around 0.80) below those advocated for less restrictive models, such as total or partial aggregation models (>0.90) (Bagozzi & Heatherton 1994). This model assumed the AQoL dimensions were independent (thus it assumed no correlations between the first level dimension disturbances), and that for each item any common variance was explained by one latent factor only. Analysis of the model, based on correlation and regression weights analysis, confirmed these assumptions (Pedhazur & Schmelkin 1991; McArdle 1996). Under these circumstances the loadings within the model also represent the correlations between the model components.

Figure 4 shows that, on average, the correlations between the five latent dimensions and the manifest items averaged 0.64 explaining an average of 41% of the item variance. The loadings of the five first order latent dimensions on the generic HRQoL index were 0.64 for the Illness scale (explaining 41% of the variance within the Illness scale), 0.67 for the Independent Living scale (45% of scale variance), 0.77 for the Social Relationships scale (59% of scale variance), 0.51 for the Physical Senses scale (26% of scale variance), and 0.87 with the Psychological State scale (76% of

scale variance). The overall comparative fit index (CFI) was 0.90, indicating a much better fit than might be expected under the restrictive conditions of model construction outlined above (Pedhazur & Schmelkin 1991; Bagozzi & Heatherton 1994).

Figure 4 **Structural equation analysis**



Summarizing these results, the analysis indicated that 90% of observed variation between observations may be explained by the structure of the AQoL. There is virtually no addition to explanatory power through relationships not postulated by the model.

Some preliminary evidence is available regarding criterion (concurrent) validity, where the criteria were other independent measures. Three such measures, each measuring an important aspect of HRQoL, are presented here: a measure of functional status (the Barthel Index (Wylie & White 1964; Mahoney & Barthel 1965)) a measure of mood (the Affects Balance Scale (ABS) (Derogatis 1992)), and a measure of general health (the SF-36 (Ware et al 1993)). Since each of these instruments measures a different aspect of HRQoL, moderate correlations — $r = 0.40-0.70$ — between the AQoL and each instrument were expected. Figures 5, 6 & 7 show the regression line and 95% CIs between the AQoL and each of these instruments' scores.

Figure 5 **AQoL/Barthel index**

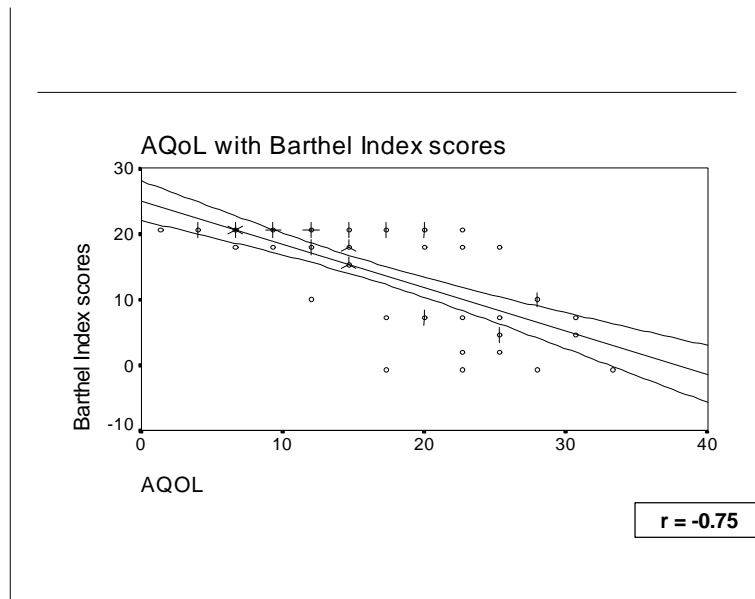


Figure 6 **AQoL/Affect balance scale**

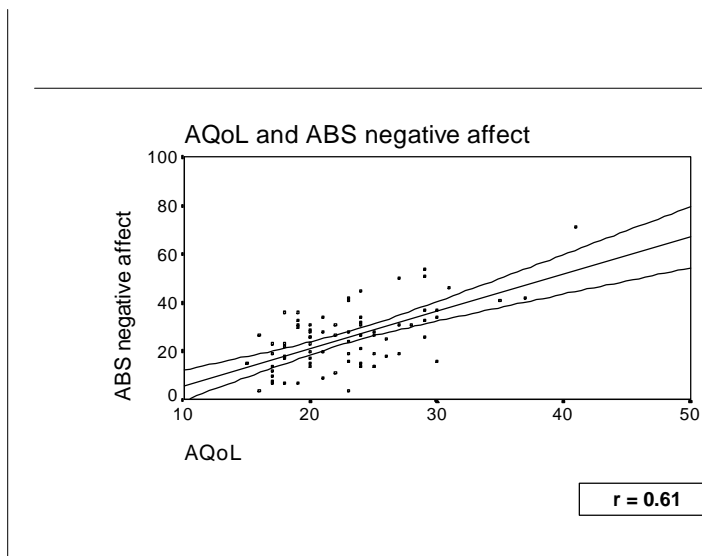
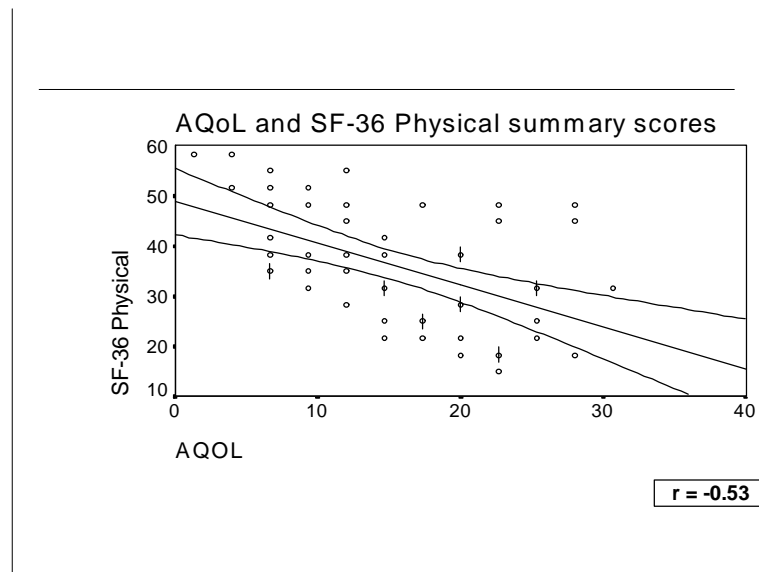


Figure 7 **AQoL/SF-36**



The data in Figure 6 are from 80 people attending a stress management program, and the data in Figures 5 & 9 are from 60 stroke victims assessed by a clinician at three and six months afterwards.

As shown in the three figures, the correlations between AQoL scores and the criterion scores were as expected, given the different instruments were tapping into different aspects of HRQoL. The figures would suggest that the AQoL is sensitive to different affective state levels, that it is sensitive to different levels of functional capability, and it discriminates between those with different levels of general health

1.3 Utilities

The validity of MAU-instruments depends upon the achievement of structural² and preference (utility) independence³. The latter is usually assumed or achieved by careful item selection (Winterfeldt & Edwards 1986; Torrance 1986). The AQoL achieves the former through the orthogonality of its dimensions, as described above (see Figures 2 & 4).

The character of an MAU instrument will reflect three key decisions; viz, (1) which scaling method is used (standard gamble, time tradeoff, etc); (2) what form of model is employed to combine scores (additive, multiplicative or statistical interpolation from the values of a limited number of health states); and (3) the relationship between the initial model scores and utility scores on a life-death scale (direct utility measurement

² Structural independence means that the dimensions are not statistically correlated. This is a required property to prevent the 'double-counting' of some aspect of HRQoL. For example, if the frequency of pain was correlated with disruption to the activities of daily living caused by pain then the inclusion of items measuring both of these attributes could result in the double counting of the disutility of pain.

³ Preference independence is where the value assigned to a response on one item does not depend upon the level of another item. For example, if the disutility of deafness depended upon the quality of personal relationships then there would not be preference independence between these two dimensions Feeny, Furlong & Torrance (1996).

or linkage between the model and a life death scale). For reasons discussed by Richardson (1994) and Dolan et al (1994), scaling was carried out using the time tradeoff technique. Interviews were conducted with a random sample of 350 Victorians within electoral divisions stratified in order of selected characteristics of the Australian population. Respondents were asked to evaluate each item response on an 'item best-worst' response scale; the item worst response was evaluated on a 'dimension best-worst' scale; dimension all worst health states and the instrument all worst health state were measured on a 'good health-death' scale.

The hierarchical model of the AQoL shown in Figure 4 reduced the (inevitable) tradeoff between instrument sensitivity and the need for response orthogonality between dimensions. Within dimensions there was no attempt at achieving structural independence, thereby allowing greater descriptive accuracy. The possibility of double counting was overcome by limiting the possible disutility from each dimension to the disutility of the dimension all worst health state as independently measured.

Since the AQoL represents approximately 1.6 million health states (4^{12} combinations of responses)⁴ direct utility measurement of each state is impossible. Of the two feasible MAU models available—additive and multiplicative—the latter is significantly more flexible and was adopted for modelling each of the five dimensions and the overall instrument score. This model creates a score between 100–0, which must be then recalibrated on a 'good health-death' scale. This was carried out using the single value of the instrument 'all worst health' state on a 'full health-death' scale. If this value is incorrect then all of the MAU values will be systematically biased. When this approach was used with the AQoL a result was obtained which (as elsewhere) predicted lower MAU scores than those directly observed. As a consequence the final utility scores were computed from the four (independent) utility values of the dimensions' 'all worst health' states. While this resulted in significantly higher utility values these have not, to date, been independently validated.⁵

Dis-utility value (DU) results for the 15 items are given in Figure 8, where these were constrained between 0–1. These may be inserted in the first five multiplicative equations presented in Figure 9—the dimension formulae—to estimate an index number for each dimension's utility (on a 100–0 scale, where 100 and 0 represent the index number for the dimension 'all worst' and 'all best' respectively).

⁴ For logical reasons the illness dimension is not used in computing utility scores.

⁵ This situation is not unique: no utility instruments have been reported in the literature as having been independently validated in this way.

Figure 8 Item dis-utility values

Dimension	Item	Health level			
		1	2	3	4
Illness	1.	0.000	0.328	0.534	1.000
	2.	0.000	0.269	0.467	1.000
	3.	0.000	0.166	0.440	1.000
Independent Living	4.	0.000	0.154	0.403	1.000
	5.	0.000	0.244	0.343	1.000
Social Relationships	6.	0.000	0.326	0.415	1.000
	7.	0.000	0.169	0.396	1.000
Physical Senses	8.	0.000	0.095	0.191	1.000
	9.	0.000	0.147	0.297	1.000
Psychological Wellbeing	10.	0.000	0.145	0.288	1.000
	11.	0.000	0.253	0.478	1.000
	12.	0.000	0.219	0.343	1.000
	13.	0.000	0.107	0.109	1.000
	14.	0.000	0.141	0.199	1.000
	15.	0.000	0.104	0.312	1.000

Figure 9 AQoL utility formulae

$$DU1 = (1.1641 * (1 - (1 - 0.3350 * U1)) * (1 - 0.5927 * U2)) * (1 - 0.4896 * U3))$$

Equation 1: Illness

$$DU2 = (1.0989 * (1 - (1 - 0.6097 * U4)) * (1 - 0.4641 * U5)) * (1 - 0.5733 * U6))$$

Equation 2: Independent Living

$$DU3 = (1.0395 * (1 - (1 - 0.7023 * U7)) * (1 - 0.6253 * U8)) * (1 - 0.6638 * U9))$$

Equation 3: Social Relationships

$$DU4 = (1.6556 * (1 - (1 - 0.2476 * U10)) * (1 - 0.2054 * U11)) * (1 - 0.3382 * U12))$$

Equation 4: Physical Senses

$$DU5 = (1.2920 * (1 - (1 - 0.1703 * U13)) * (1 - 0.2554 * U14)) * (1 - 0.6347 * U15))$$

Equation 5: Psychological Wellbeing

$$U = ((1.04 * ((1 - (0.613 * 0)) * (1 - (0.841 * DU2)) * (1 - (0.855 * DU3)) * (1 - (0.931 * DU4)) * (1 - (0.997 * DU5)))) - 0.04)$$

Equation 6: AQoL utility score

For each dimension there are 64 possible health states, the disutility value of which may be estimated either directly from equations 1–5, (where the values of $U1$ and $U2$ are the item disutility values in Figure 8) or from the disutility values in the look-up tables provided in Appendix II. Note that ‘utility’ scores may be derived from the disutility scores using the formula:

$$U = 1 - DU$$

Formula 1

where D and DU are the dimension utility and disutility values respectively. They represent ‘utility’ scores as they were derived using the TTO procedures. However

the scores are calibrated so that for each dimension the ‘dimension all best and ‘dimension all worst’ utility scores take values of 1.00 and 0.00 respectively. This means that ‘utilities’ from different dimensions cannot be directly compared (as dimension ‘all worst’ health states do not have the same utility values on a life-death scale). For the same reason dimension utilities cannot be compared with utilities measured on a conventional scale where 1.00 and 0.00 represent full health and death respectively.

The disutility values from four of the five dimensions may be combined to produce an overall utility score using Equation 6. The first dimension, Illness, is not used in the computation of utilities. This is because there was a logical problem with respect to it. It had been constructed to reflect an underlying pathological health state. Instead of measuring this directly, we determined to measure it via indicators of health service consumption on the assumption that these would reflect the level of the pathological health state. The difficulty with this approach is that when a person consumes health services (e.g. medication, treatment or use of medical aids) the therapeutic effect of this is lost. Therefore this dimension was omitted from the utility computation. The four dimensions used in utility computation are Independent Living, Social Relationships, Physical Senses and Psychological Wellbeing.⁶

The numbers obtained from the equations refer to utility values on a conventional scale where 1.0 and 0.0 refer to good health and death respectively.

Consistent with the Torrance (1986) HRQoL-health states worse than death can be measured by the AQL. Torrance argued for an arbitrary lower endpoint of -1.0 .⁷ Although this was used as the theoretical lower boundary when eliciting preferences, the range of utilities for the AQL ‘all worst health state’ scenario on the ‘full health – death’ scale is -0.04 — $+1.00$. The meaning of the negative scores associated with health states worse than death is as follows. A score of, for example, -0.03 for a particular health state, indicates that a person would regard death as being equally (un)attractive as being in full health for 3% of future years and in health state A (the comparator health state; ‘full health’) for the remaining 97% of future years.⁸ That is, a negative health state implies that a person requires some time in a full health state to compensate them for this option to be equivalent to death.

Regarding the reliability of the AQL-utility values. The internal consistency for the construction sample’s TTO-values was $\alpha = 0.86$.

⁶ This does not imply that the Illness dimension has no value; it provides an estimate of the value of health care resources consumed. Consequently it is included in the discussion here.

⁷ Preference scores are obtained in a particular context. In the case of the TTO, where positive values are reported this context is the trading-off of years of poor HRQoL from years of ‘full health’. The inferences here are reasonable. By contrast, when TTO values are in the negative range the TTO question requires people to determine that the relationship between ‘full health’ and the health state to be evaluated is such that this relationship is equally desirable as death. This is an unrealistic context and our confidence in the meaning of the preference that is inferred here is correspondingly uncertain: the more extreme negative values cannot be interpreted literally. For a full discussion of this issue, readers should consult Richardson & Hawthorne (2000) Negative Utility Scores and Evaluation of the AQL All Worst Health State, CHPE, WP73.

⁸ Mathematically this can be expressed as: $0.03U + 0.97 \cdot (1.00) = 0.00$. Solving this gives:

$$U = \frac{-0.03}{0.97} = -0.03.$$

2 Instrument administration

There are several issues which have come to our notice regarding the administration of the AQoL and the coding of AQoL data. This section outlines these issues and discusses how we suggest they should be handled.

In general, the AQoL has proven to be easy to administer, to understand and to complete. The reading level of the AQoL is 71% on the Flesch Reading Ease scale, suggesting its reading level is appropriate for most literate people. Experience with interview administration indicates that it takes about 5-7 minutes to complete. Respondents have also reported it is easy to approach, comprehend and complete.

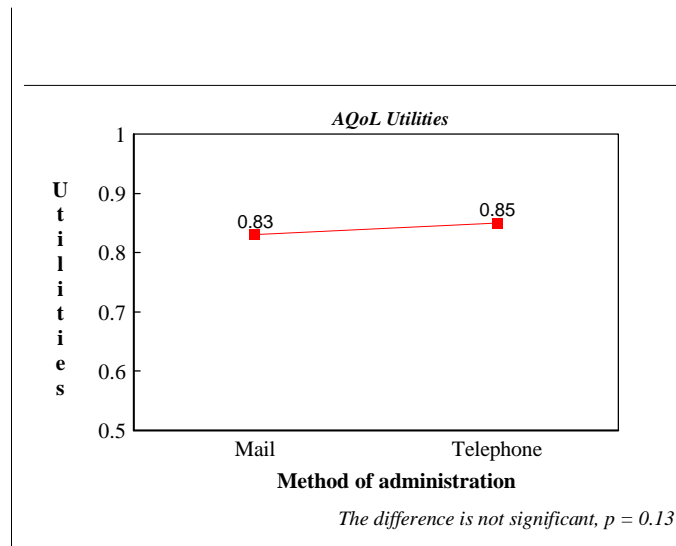
2.1 Mode of administration

The AQoL was designed for self-administration on the assumption that it would primarily be used in field settings where trial participants were able to self-complete: receiving the AQoL either through the mail or being handed it during an interview. This is consistent with Guyatt et al's (1996) findings that self-administration not only increases the response rate but also leads to lower rates of missing data. For this reason, and to personalise the perspective, the items were all written in the first person.

There are, however, situations where the AQoL will be used where other methods of administration will be employed. Typically, these will be either over the telephone, by direct interview or completion by proxy (e.g. by a caregiver).

Most commonly, we expect researchers will be interested in the effect of telephone vs. mail administration. To investigate this, we randomly sampled 139 cases from a health status population survey and asked them to complete the AQoL both by telephone and by mail. Participants were randomly assigned to each method in turn. The interval between administrations was timed to be two weeks: long enough for memory to play little part in the results, yet close enough so that respondents' lives would not have dramatically changed. The results are given in Figure 10 which shows that there was no significant difference in the mean utility score. The Pearson correlation between mail and telephone overall AQoL-utilities was $r = 0.66$. This finding suggest that researchers should not use mixed methods of data collection.

Figure 10 Mail vs. telephone administration



2.2 The time frame

Although no time frame was included within each item, the introductory instructions asked respondents to evaluate their health state over the *previous week*.⁹ Thus items include comments like ‘generally’, or ‘only occasionally’. This was an attempt to overcome the difficulty that respondents’ HRQoL health status—and their assessment of that status—may vary quite considerably within a very short timeframe. Where the timeframe is a long one they may be affected by real health status changes, by the use of drugs, or by recall bias. For example, pain is often episodic. Asking for assessment of their health state *now* or *in the past* may provide misleading assessments.

A key point to note is that this evaluation refers to how the respondent **was** during the past week: it does not ask respondents to assess health state changes within the past week nor to assess the degree of change compared with a previous timeframe. If respondents are confused about this, they should be informed that the AQoL questions ask about their health state during the past week: it is assumed this is a global overall evaluation for that period.

2.3 Items which may cause difficulty

The AQoL has been administered to thousands of respondents with very few reported difficulties. Generally respondents and interviewers have commented that it is easy to understand and to complete. Some of the items, however, may cause some respondents to seek clarification. In general, about 2% of respondents will have a question about the instrument or an item. The items about which questions have been asked are discussed.

⁹ This does not, of course, imply that utility scores only apply to a week. The items in the instrument are used to gain descriptive information about health states. The value of these health states is subsequently inferred from interviews in which the stated duration of the health state is usually much longer. The TTO-values used in the AQoL were obtained using a 10-year timeframe.

Items 1, 2 & 3 Problem of confusing treatment with the outcome. The treatment being evaluated must be excluded from the response. For example, if the treatment consists of a new drug (Drug Q), when respondents in the treatment group are asked to complete Item 1, *Concerning my use of prescribed medicines* and they include Drug Q this may cause them to select a response lower (i.e. to report higher medication use) than they would otherwise select. Respondents should be informed that their responses to these questions should not include the treatment of interest.

Item 2 This item refers to “rely(ing) on medicines or a medical aid”. Two issues have been noted by AQoL users: it is a double-barrelled item probing either reliance upon medicines and/or aids; and it is repetitious of Item 1 which probes the number of prescribed medicines used.

‘Medicine’, as defined during the construction of the AQoL, refers to any remedial substance used to restore or maintain health; a medical aid is similarly defined here as any mechanical or electronic aid used for the same purpose. This would, for example, include a back brace as well as a heart pacemaker; the only aids excluded are glasses or a hearing aid. These were excluded on the grounds that there are explicit questions covering these later in the AQoL; the exclusion avoided double-counting. As such, this item refers to reliance upon the broader family of medicines or medical aids; many of which may not be prescribed by a registered doctor but may be provided by an allied health professional. This may be contrasted with Item 1 which explicitly probes the use of prescribed medicines.

Some respondents also may think that ‘medication’ here refers only to liquid medications; it refers to medications of all kinds, including lotions, liquids and solids.

Item 6 Some respondents have questioned the change in item responses 2 & 3. Response 2 lists getting around the home before getting around the community; response 3 reverses this order.

This change in word order was determined during item trialing when respondents reported that they found it difficult to differentiate between the two levels and tended to respond at random. By reversing the order, it was found that respondents thought more carefully about the issue.

Item 7 This item may cause some respondents some distress for it asks them to reflect upon their intimate personal relationships. We have encountered some people who did not wish to answer this item because it made them feel anxious, worried or depressed.

-
- Item 8** See Item 7. The responses to this question are double-barrelled and a person who chooses to be socially isolated or who is socially isolated because of illness may not necessarily be lonely. Where this is the case, the first part of the response should have priority; i.e. a person who is socially isolated but not lonely, would be advised to select option 4 if they asked for clarification.
- Item 13** Some respondents have asked for an intermediate response category between responses 3 & 4 on the grounds that they wake up at night and are usually unable to go back to sleep.
- Logically, these respondents fit response category 4. If they wake up at night and are unable to go back to sleep, then they are awake most of the night. To investigate if this was true, the item was administered in two forms to a sample of 63 cases drawn at random from the generally community and a Melbourne hospital. One form asked the item with the 4-point response scale and the other used a 7-point scale, with interpolated values. Only one case used the interpolated value between responses 3 & 4.
- Item 14** Some respondents have misunderstood the use of the word ‘or’ on the responses, and have assumed they have to suffer from all three afflictions. The responses indicate that a person only needs to suffer from one of these conditions to respond beyond level 1.
- Item 15** Occasionally a person will ask how to gauge how much pain he/she suffers from. He/she should be advised that this is an internal assessment; i.e. his/her own evaluation. We have received one or two enquiries asking whether responses to this should reflect pain status without medication when on medication for pain relief (e.g. use of morphine). The advice should be that this question is to be answered for his/her current situation, irrespective of medication use. Likewise, if a person avoids certain activities because they cause pain, where those activities are avoided the person is not currently in pain.

2.4 Missing data

There are several reasons for missing data occurring in a study. These may be related to the study, or independent of the study. These reasons include: a simple skipping of a page or question (either deliberately or accidentally); a respondent may miss an appointment; they may refuse to complete the AQoL; program staff may forget to administer the AQoL; there may be language or reading difficulties; the respondent may no longer be ill — i.e. they may have recovered, either naturally or because of the treatment —, or they may be too ill (which may be determined by a nurse rather than the patient themselves); they may have died, perhaps naturally or from treatment related causes; some respondents may move home; or they may simply wish to leave the study (e.g. they grow tired of participating).

Grouping these reasons, generally there are two kinds of missing data:

-
- Monotone or censored missing data. This is where a study drop-out is permanent. This may be caused by withdrawal or death. Where a chronic disease is involved, it may be because the patient has been cured or because the treatment doesn't work. It is likely with progressive fatal diseases, such as advanced cancer or AIDS.
 - Intermittent missing data. This is where the missing data is non-permanent. E.g. where a visit is missed, but the subject comes back in the study later. Maybe it is where a question of the AQoL is missed, or a whole page.

Generally there are two possible effects of missing data on a study:

1. There will be fewer observations leading to a loss of study power; and/or
2. Where the missing data are non-random, this will lead to bias in study findings. Where subjects are randomized into groups and some drop out of the statistical analysis, this may induce bias which favors one group (Gould 1980).

Consequently it is important that every effort is made to avoid missing data during the collection phase.

Monotone missing data

There is little a researcher can do if faced with substantial monotone AQoL missing data. The two traditional ways of handling this situation are to either exclude the withdrawals or to conduct an endpoint analysis. Where the timeframe is reasonable, endpoint analysis is to be preferred.

Exclusion Here the analysis would only include those respondents remaining through to the end of the study. I.e. this allows the researcher to make statements about respondents who remained on the study for part of the full treatment period.

The limitations of this are twofold:

- If the rates of withdrawal between the two groups are different, this will lead to spurious results; and
- If cases withdraw for treatment-related reasons, those remaining in the study may no longer be representative of the population.

Endpoint analysis This is where drop-out cases' last recorded AQoL values are used for the final AQoL measurement; i.e. respondents' utilities are carried forward. This avoids the problem of non-comparable samples, but gives widely differing data (by time, length of treatment etc). It also assumes that:

-
- The endpoint data would be the 'true' data had the case remained in the study until the end; and
 - There is no distinction in the analysis between true 'observed' data and 'assumed' data.

Intermittent missing data

Based on our experience with the AQoL, an intermittent missing data rate of about 1% per question can be expected in mail surveys.

There are three kinds of intermittent missing data, each of which has different implications for the AQoL findings. They are:

1. Missing completely at random (MCAR). This is where missing data are independent of the AQoL instrument. This includes the respondent moving address,¹⁰ the staff forgot to administer the AQoL (in both these cases the AQoL is missed at, say, baseline, but it may be administered at a follow-up), or an item or page of the AQoL was inadvertently missed. The effect on the study is to diminish the study's power to detect real differences due to the treatment (if they exist).
2. Missing at random (MAR). This is where data are missing independently of the current AQoL data collection, but where they may be explained by previous AQoL or other data collection. This may arise, for example, where there was an unfavorable reaction to previous data collection (e.g. a dislike of the interviewer; or completion of the AQoL causing distress which was not conveyed to the researchers). I.e. MAR involves events which cannot be predicted; their effect on the study findings is uncertain.
3. Not missing at random (NMAR). This is where data are systematically missing for an observable reason. For example, during construction of the AQoL it was noted that 21% of respondents refused to answer questions probing their sexual behaviours; analysis showed these data were missing for particular types of people. Since these were NMAR data which could have biased AQoL scores, these questions were excluded during the construction phase of the AQoL. Examples which may occur in studies using the AQoL could include the death of patients due to their disease, the toxicity of treatment, refusal to answer specific questions etc. Note that these missing data are dependent upon the real current value of the AQoL (e.g. if a person dies their current real utility is '0.0'; if they have been returned to good health it is '1.0'). If there is any reason to assume NMAR, then missing data must be accounted for during data analysis as its omission will bias the study findings (Fairclough 1998).

When handling missing data, it is important to determine which kind of missing data is being reported, and to take whatever steps are appropriate to collect it.

¹⁰ If there is any reason to suspect that the reason for moving is related to the respondent's socio-demographic characteristics, then moving address may not be strictly MCAR.

Often MCAR data can be collected through adequate follow-up reminders (e.g. telephone or interview follow-up to a mail AQoL administration; several follow-ups may be needed). Alternatively, if the MCAR data are a single page or a single item missed, it may be possible to photocopy the AQoL page and forward it to the respondent with a polite note asking him/her to complete it.

For MAR data, it is important to try and identify the cause of the problem. Once this cause has been identified, remedial steps can often be taken to minimize MAR data. For example, inadequate training of an interviewer may cause MAR data. If this can be identified through appropriate audit procedures (e.g. by monitoring interviewer/data relationships), further training can be provided to the interviewer.

If NMAR data are observed, steps should be taken to determine the direction of the bias on the study's findings and this reported along with the analyses. It may be possible to adjust analyses for an identified bias. Consideration should be given to endpoint analysis.

For MCAR and MAR data which remain missing after taking all reasonable collection steps, where this is confined to individual AQoL items, it may be possible to determine the values using one of the following methods. Generally, the most appropriate method will be to model respondents' data. If this cannot be done, the value may be imputed using either of the procedures outlined. The steps taken to handle missing data should always be fully reported.

Modelling values

This involves careful statistical examination of the database in order to model the missing values (usually through use of a regression equation). The procedure is to identify the independent characteristics of all respondents who have responded to the item for which the datum is missing. The characteristics of the missing datum-respondent can then be matched with those who responded. Once a match is found, the average value for the matched cases can be applied.

For example, suppose Respondent A provided these data: Q1: 2; Q2: 2; Q3: X. The characteristics of A (female, >70 years, breast cancer, etc) were matched with those of all other similar cases (female, >70 years, breast cancer, etc). Examination showed that the mean score for Q3: 3. This value would then be assigned to A.

The two advantages of modelling are that the assigned values will not bias the findings nor will they artificially reduce the variance within the database.

The disadvantage of modelling is that it requires a large database and careful analysis. It also requires that the number and type of independent variables collected are carefully thought through prior to any data collection.

Imputing values I: Averaging within cases. This refers to computing each case's average item score within each dimension where there is a missing datum, and then assigning this average score to the missing item. The advantage is that it is simple to do and it reflects the individual's values for the dimension in which the datum is missing.

The disadvantage is that the averaged values may reduce the variance within an AQoL dimension, and thereby bias the true utility value associated with the real condition. This problem becomes even more acute if data are missing from more than one item within a dimension.

For example, consider an individual who responded: Q13: 1; Q14: X; Q15: 3. The averaged response level for Q14 would be $1 + 3 = 4/2 = 2$.

Imputing values II: Averaging across cases. Missing data can be imputed by calculating the average score across all cases and applying this to the item for which the datum is missing. Generally, this method is not recommended and should only be used as a last resort after all else has failed.

For example, for Respondent B, Q9: X. The mean for Q9 across all cases was '1'. This would be assigned to Respondent B, thus Q: 1. The advantage is that this procedure is quick and easy.

The first problem with this procedure is that it imputes a value which may be completely inappropriate (B's responses to the other questions in the dimension might have been Q7: 4; Q8: 4; Q9: X. To impute a value of '1' would probably be highly misleading). The other disadvantage is that it will artificially reduce the variance within an AQoL dimension, and thereby bias the true utility value associated with the real condition. This problem becomes even more acute if data are missing from more than one item within a dimension.

Missing values and the computer algorithm

The computer program in Appendix III contains several lines of computer code which impute values, using the averaging within case approach. These lines will automatically detect when a value is missing from one variable within a dimension and impute the value to be the mean integer value of the two other variables within that dimension. For example, if the data were: Q10: X, Q11: 1, Q12: 3. The program would impute the value for Q10 as $(1+3)/2 = 2$.

Where more than two variables are missing from within any dimension, the research team would have to use one of the techniques above. It is recommended that where

intermittent missing data comprise 30% or more of all responses the case should be discarded.

Note: Missing data are handled by imputing values within each dimension. Regression analysis should not be used to impute values across dimensions because this may violate the statistical independence of each dimension and result in double-counting.

2.5 Using proxies

Although the AQoL was designed for individual self-completion, there will be times when individuals cannot meaningfully complete it. At these times it may be appropriate for others to complete it; this may be preferable to excluding these cases from a study (Sneeuw et al 1997a).

Determining who the proxy should be may be problematic: caseworkers, doctors, spouses, parents or children, and friends may all be available. The most appropriate proxy, however, is the respondent's main caregiver. This caregiver is likely to know their caree's real situation and be able to represent it.

Because there are difficulties with mixing proxy and respondent scores in the one database, the literature on proxy completion of HRQoL-instruments should be consulted. Within this literature there is, generally, considerable variation in the methods used, the selection of appropriate proxies, the procedures recommended and the conclusions drawn. For a review see Sprangers & Aaronson (1992). The substantive conclusions would appear to be that proxies consistently rate the respondent's HRQoL lower than does the respondent him/herself (although this is not the case for all dimensions of HRQoL) (Sneeuw et al 1997b); and clinicians seem particularly prone to misrating pain and social relationship levels. We have found that proxies' scores are consistently lower than respondents, and that this is particularly evident on the Social Relationship dimension (Hawthorne, Herrman & Thomas 2000).

There is also general agreement that a proxy's ratings are influenced by his/her own health state, particularly where there is an onerous caregiver burden. Other factors affecting a caregiver's responses include a false consciousness of the true state of affairs, differences in perception and valuing, and that there may be mutual dependence (Raphael 1996).

Where proxies are used, it is strongly recommended that a stratified analysis be undertaken in order to determine the impact of the proxies' scores on the study findings. Where such an analysis demonstrates that proxies' scores make a major contribution to the study findings, it is recommended that adjusted proxies' scores are computed and used in the data analysis.

2.6 Sample size

One of the most commonly asked questions relates to the sample size required by researchers to demonstrate differences as a result of an intervention. Since the AQoL is a new instrument, many of its properties are not yet known, including its sensitivity under different circumstances and with different populations.

Consequently no definitive sample size estimates can be given; the sample size needed for any particular study should be worked out by the researchers who have expertise in their particular field. It is recommended that this process involves a statistician with whom the researchers can directly liaise.

Subject to this caveat the following advice may be found useful.

Issues in sample size calculation

Sample size calculation is concerned with including sufficient subjects in a study to avoid Type I and Type II errors.¹¹ Sample size calculations only ever specify the minimum number of subjects for a study: confounding requiring stratification during analysis is invariably ignored and requires greater numbers of subjects. Also, as the n increases, the standard deviation also approaches the true population standard deviation. Several things follow from this:

- As the sample size increases the variance will diminish (i.e. the sd or $95\%CI$ will become narrower, hence the argument to 'increase the sample size' to obtain a significant result). However, since a distribution resembles a normal distribution beyond about 30 cases, if a pilot study has >30 cases it can be assumed that increasing the sample size will not make huge differences to the sd or $95\%CI$ for calculating the main study sample size.
- Where a study involves several criteria, sample sizes should be calculated for each criteria. If the criteria are of equal importance, the largest calculated sample should become that used in the study. If one criterion is the 'gold standard', the sample size for this should be used.
- It also follows that with a large enough sample size the null hypothesis will always be rejected.

The issue for sample size calculation is to have a large enough sample to demonstrate a true effect if there is one, but small enough to avoid rejecting the null hypothesis when it should be accepted. When calculating sample size there are five key issues:

- The 'size' of a test (α);
- The 'power' of a test ($1-\beta$);
- The 'effect' of the variable of interest (δ);
- The 'variability' of the variable (sd); and
- The 'unit of analysis'.

¹¹ However, it should be noted that sample size calculation does not demarcate between studies worth pursuing and those not worthy.

The size of a test (α)

This is the probability of the researcher making a Type I error (α ; i.e. concluding there is a difference when there really is no difference). The probability of a Type I error should be set by the researcher at the research design stage. The conventional level is $p = 0.05$, i.e. there is a 5% chance of committing a Type I error.

The power of a test ($1-\beta$)

The power of a study is the probability of the researcher making a Type II error ($1-\beta$; i.e. concluding that there is no significant program effect when there really is a program effect). The lower the power the less certain a researcher can be of the results. The power of a test depends on the size of the effect a researcher wishes to detect and the consequent determined sample size needed to detect this effect. High power requires valid, reliable and sensitive instruments especially where a large sample is not possible. In either case, however, the calculation of power is a necessary prerequisite for the calculation of the required sample size. In medical research the convention is for power to be set at 0.80. In evaluation and educational research — where the effect size is usually small (see below) — the power of a study is often set at 0.90.

The effect of the variable of interest

When a new program or treatment is introduced it will be expected to have a greater effect than the program or treatment with which it is being compared. The unit of measurement used is delta (δ or d). This can be thought of as the effect size of the treatment (also referred to as the *critical effect size* [Δ] when the question is asked: how small does the increase in a score have to be to be of importance? Thus Δ describes the *determined effect size* while δ refers to the *obtained effect size* (Kraemer & Thiemann 1987). δ is one standard deviation from the mean of an obtained score, i.e. one z-score unit. In general the effect size for non-paired data can be calculated by the formula:

$$d = \frac{\bar{X}_e - \bar{X}_{ne}}{sd_{ne}}$$

Formula 2

Where:

\bar{X}_e = the mean score of the experimental group,

\bar{X}_{ne} = the mean of the comparator group, and

sd_{ne} = the standard deviation of the control group. In a pre-post test the sd is the pooled standard deviation (ie sd_{pooled}) (Cohen 1988).

In the medical field the accepted interpretation of δ is that:

- 0.00 – 0.29 = a small effect

-
- 0.30 – 0.49 = a moderate effect
 - 0.50 – 0.69 = a good effect
 - 0.70+ = a large effect (Cohen 1988).

The accepted effect size for a study should take into consideration the practical real-world values of findings. Extremely small effect sizes, although statistically significant, may not justify the effort of finding them, or, if they are found, have a great deal of practical value. It is prudent, therefore, to adopt realistic effect size values.

The variability of the variable (sd)

Where two groups are to be compared, this is simply the standard deviation of the control group. For paired data (e.g. pre-treatment and post-treatment AQoL scores) the pooled standard deviation can be used. It can be estimated from:

- A pilot study
- Previously published studies
- An informed estimate (e.g. if the AQoL mean score was expected to be 0.75, then the informed estimate for the *sd* might be 0.08 based on a guess of 10%).

The unit of analysis

Where programs are administered to groups of people (e.g. respondents from two hospitals, or from three communities), it has been argued that the unit of analysis should be the group rather than the individual.¹² Statistical tests, such as independent t-tests and analysis of variance are based on the assumption that the units of data used for the analysis are independent. Subjects in groups may not necessarily be independent of each other; they may interrelate in some way such that they form clusters (e.g. where a treatment is administered to the group or where people come from separate clinics or hospitals). The effect of clustering is to reinforce common knowledge, attitudes and behaviors. Under these conditions the statistical assumption of independence of the error term is violated leading to a Type I error (i.e. concluding there is a program effect when there is actually no effect). Consequently it has been argued that group means should be used in order to avoid violating the independence assumption underlying many statistical tests (McCullagh & Nelder 1983). This clustering effect can be controlled for by increasing the sample sizes above the minimum required. Where it is likely clustering occurs, researchers should consult with a statistician and the relevant literature on intra-cluster correlation for computing sample sizes.

Sample sizes for the AqoL

¹² This does not apply to paired data, where individuals may act as their own controls through providing pre- and post-treatment scores. It does apply, however, whenever there is a control group with which the treatment group's scores are being compared.

Sample size estimates under a variety of assumptions are presented in Figure 11. These estimates are based on the following parameters, which are those most commonly experienced by researchers. It is assumed the study design comprises cases randomized into treatment and comparator groups of equal size, and that post-treatment AQoL scores are to be compared (i.e. use of an independent or dependent t-test or Mann-Whitney U-test, depending on the data distribution).

- The 'size' of the test (α) has been set at $p = 0.05$.
- The 'power' of the test ($1-\beta$) has been set at 0.80.
- The 'effect' (δ) of the variable of interest is presented for a range of estimates; from 0.10 through to 1.00.
- The 'unit of analysis' is assumed to be the individual case.

If the assumption is made that $sd_e = sd_{ne}$ (which usually implies that the two samples are similar sized¹³), then the generic sample size formula for estimating the n in each group can be used:

$$n = \frac{2sd^2 \cdot (Z_a + Z_{1-b})^2}{(m_e - m_{ne})^2}$$

Formula 3

This can be simplified (Formula 10) where the effect size is predetermined, as in Figure 11:

$$n = \frac{2 \cdot (Z_a + Z_{1-b})^2}{d^2}$$

Formula 4

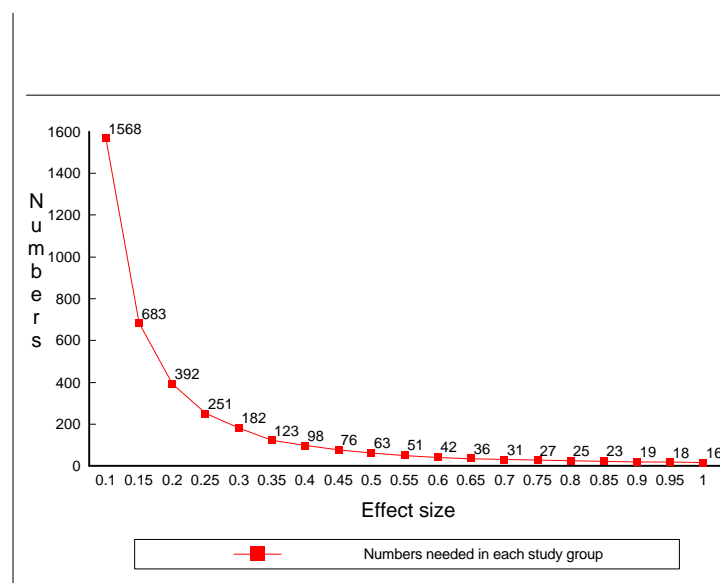
- sd is estimated as described above.
- Z_a is for a two-sided test, with $p = 0.05$, $z = 1.96$
- Z_{1-b} is for set for 80% power, thus $1 - b = 0.20$, $z = 0.84$
- $m_e - m_{ne}$ in Formula 3 is the specified difference between population means. This should be determined by the smallest estimated clinically significant difference

¹³ This is usually true; there may, however, be situations where the numbers in each group are equal but the variance within the groups different.

- ***d*** is the predetermined effect size of interest. This should be determined by the smallest estimated clinically significant difference which is important; i.e. for sample size calculation it should be assumed that $\Delta = d$.

When using Figure 11 to estimate minimum sample sizes, once an estimate of ***d*** has been made, the corresponding number of subjects required for each cohort in the study can be read off on the Y-axis. For example, if in a study of a new drug for blood pressure, the Δ was estimated to be 0.25 and it was assumed that $\Delta = d$, the numbers in each group of the study would need to be 251, i.e. 502 subjects would need to be recruited into the study and then randomly assigned to treatment/comparator groups.

Figure 11 **Sample size**



2.7 Languages other than English

The AQoL has not been translated into any languages other than English.

Where a non-English-speaking person is to be administered the AQoL, it is recommended a professional interpreter is employed. The quality of the translation both from English and back into English cannot be assumed where an untrained interpreter (e.g. family member) is used given the difficulties of achieving accurate concept and language equivalence (Orley & Kuyken 1994).

2.8 Changing AQoL items

There may be times when researchers wish to alter AQoL items.

The AQoL was designed to be used as a generic instrument applicable across a wide variety of healthcare interventions where the utility values would enable comparisons of interventions and/or economic evaluation. As such it is important that the results from different studies can be directly compared. The validity of such comparisons is determined by the use of a common instrument. We recognise, however, that all interested AQoL users reserve their own rights to publish information to further the

validity and reliability of the AQoL; any investigator who wishes to publish such a finding is asked to submit a copy of their manuscript to the AQoL development team for review. Where such a review indicates that the AQoL scoring algorithm has been modified, researchers should not refer to their scores as 'AQoL utility' scores.

So that researchers can use the AQoL confident that their results are truly comparable with those obtained elsewhere, the AQoL has been copyrighted and may not be used or modified without the written consent of the authors.

This restriction is not intended to discourage use of the AQoL, rather it is in the interests of all researchers as it ensures a standard instrument is available. It enables researchers to use the AQoL in full confidence that AQoL support is available, and that their results will be directly comparable with those obtained by others.

Where a more detailed assessment of particular health-states associated with a disease is required, the AQoL should be used in conjunction with a disease-specific instrument. Bowling (1991 & 1995) provides an introduction to many disease-specific instruments.

3 Scoring the AQoL

Given the psychometric and utility properties outlined in Section 1, the AQoL can be scored in two ways, depending upon the purpose of the researcher.

3.1 Data entry and initial coding

Data coding in general

When entering data from the AQoL into a computer database, the data entry will be found easiest if the rank order of item responses are entered, thus:

- 'A' = '1'
- 'B' = '2'
- 'C' = '3'
- 'D' = '4'

I.e. the higher the numerical score, the poorer the respondent's HRQoL.

This coding scheme also has the advantage that these values can be directly used to compute either summated ratings or utilities as described in the next sections.

The data should be double-entered with range-checking, and the database cleansed. Original variables should be retained and transformed/dummy variables used in all subsequent analyses.¹⁴

¹⁴ This protects the database should coding errors be made at any stage of the analyses. Ideally, a copy of the database should be used for all analyses so that the original database is kept intact.

Handling special cases

In some cases, respondents will have decided that the response options to an item are inappropriate and do not describe their situation accurately. They may either select two options, or make a written comment next to the question. Coding these cases can be difficult. Whilst different approaches can be taken to these problems, we recommend the following procedures.

- ◆ Where a person selects two responses to the one item:
 - If the two options are consecutive, the respondent should be given the benefit of the doubt and the response denoting the better health state used. E.g. if to Q11 a respondent selected 'B' and 'C', this would be coded as 'B'.
 - If the two options are non-consecutive, the mean value should be assigned. E.g. if to Q11 the respondent had selected 'A' and 'C', then a value of 'B' should be assigned.
- ◆ Where a respondent has made a written comment:
 - If they have selected a response choice this choice should be respected and used.
 - If no item response has been selected, the comment should be used as a guide for assigning a value. E.g. if to Q11 no response has been selected, but the respondent has written "I find it difficult to hear in a crowd where other people are talking", this could be assigned a 'B' on the grounds that the respondent is indicating they have "some difficulty hearing".

3.2 As an unweighted, psychometric instrument providing 'value' profiles

Where a researcher wishes to provide a HRQoL-profile, the responses to individual items can be summed. This method will provide value profiles of illness, independent living, social relationships, physical senses and psychological wellbeing. These can then be summed to provide an overall unweighted HRQoL-index.

This method of scoring the AQoL is simple and straightforward: no weighting of items is required. The disadvantage is that no claims can be made regarding utilities or preferences (i.e. the resulting scores cannot be used in economic evaluation). It should also be noted that when used in this way, the AQoL does not provide profiles of health status: the profiles are of HRQoL.

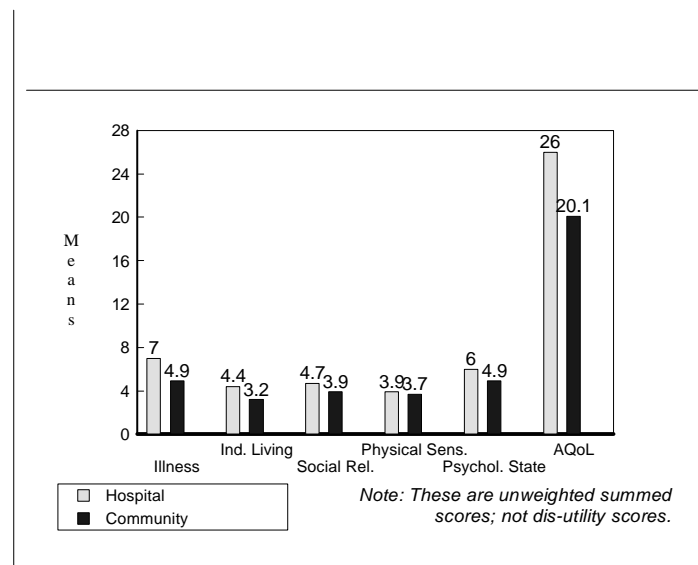
For data analyses, item responses should be recoded as follows:

- '1' = '0'; i.e. if the first response is selected, the respondent's status is 'good' and there is no loss of HRQoL.
- '2' = '1', '3' = '2' and '4' = '3'.

This coding will produce scores for each scale ranging from '0–9', where '0' represents 'good' HRQoL and '9' the worst possible HRQoL for the dimension of interest. This transformation does not alter the differences between responses, but ensures that a person with good health obtains a score '0'; if this transformation is not undertaken a person in good health will obtain a score of '3' on each scale or '15' on the summed AQoL.

After transformation, when summed into a single HRQoL-score, AQoL scores will range from '0–45', where '0' represents 'good' HRQoL and '45' the worst possible AQoL HRQoL-score.

Figure 12 Typical AQoL scores by hospital status



At the time of writing, no population norms have been established. Evidence from several studies, however, suggests that the pattern of scores presented in Figure 12 can be expected. This shows the average obtained scores from two studies involving comparison of hospital and general community cases (n = 318). These are indicative only and should not be taken as norms.

3.3 As a utility HRQoL measure

Where utility values are required, the raw AQoL data as described in Section 3.1 is used (*not* the recoded values). These raw scores will need to be transformed using the values and formulae presented in Figures 8 & 9. Since this is a cumbersome and tedious task to perform by hand, manual calculation of AQoL-utilities is not recommended. Not only is it time-consuming, but it can also lead to errors. To alleviate this problem, we have developed both 'look-up' tables for the dimensions and a data analysis program for SPSS.

In addition, the AQoL utilities can be used to calculate QALYs for use in economic evaluation.

3.3.1 Calculating utilities

Using look-up tables

The look-up tables in Appendix II (Table 1 to Table 5) provide disutility values for each of the five AQoL dimensions. Although the tables avoid using Equations 1-5, the obtained dimension utilities still have to be manually entered into Equation 6 for utility computation. This manual method may be suitable for very small data sets, although we recommend against it.

To use the look-up tables, Figures 13 & 14 should be consulted. Figure 13 provides a descriptive worked example based on the dimension shown in Figure 14, Independent Living. Figure 14 shows how to read a look-up table.

Figure 13 Look-up table: worked example

A person needs help with the more difficult personal care tasks, they need no help with household tasks, and they cannot get around the community by themselves.

- The AQoL scores:
Q4: 3 Q5: 1 Q6: 3
- The health state is described as:
3,1,3
- From Figure 15 or Table 2, the value for this health state is:
 $DU_2 = 0.4672$
- Contribution of this to his/her AQoL utility using Equation 6:

$$U_{AQoL} = 1.04(1 - 0.613^3)(1 - 0.84^1(0.4672)^3)(1 - 0.855^3 DU_1)(1 - 0.93^1 DU_2)(1 - 0.99^3 DU_3) - 0.04$$

Figure 14 Using look-up tables: independent living

		Items & health state levels			
		1	2	3	Disutilities ¹
The headings here refer to AQoL item numbers	→	3	1	1	0.2700
		3	1	2	0.4249
		3	1	3	0.4672
The entries here are the AQoL item response levels selected by respondents	→	3	1	4	0.7452
		3	2	1	0.3639
		3	2	2	0.5012
		3	2	3	0.5388
The values here describe the disutility for a person in the health state described by their item responses		3	2	4	0.7853
		3	3	1	0.4020
		3	3	2	0.5322

Data analysis by computer

This is the most efficient method of data analysis, and can be readily adapted for almost any situation. The necessary computer algorithms are provided in Appendix III. Although they have been presented in a format suitable for use by SPSS (SPSS 1996) there is no particular reason SPSS has to be used. The program has been written in a format that will work with all versions of SPSS and with any suitable SPSS data file, following the modifications described below. Although this means the program is longer and simpler in construction than strictly necessary it enables all researchers to use it easily and to modify it for use with other data analysis programs, including those in the public domain, such as EpilInfo (Dean et al 1994).

Before modifying the program, it should be typed in exactly as is, saved and run in SPSS¹⁵. If the results are identical with those provided below, the program has been correctly typed in and may be modified as described. Until these results are obtained, the program will need to be checked for errors; most likely typing errors.

Description of the program

- | | |
|---------------|--|
| Lines 33–38 | Define the dummy data to be used to test the program. This assumes the AQoL item variables are named Q1, Q2, Q3 etc. If other variable labels are used, these should be modified accordingly. This also assumes the data are in fixed field format. The data columns specified for each variable relate to the sample data set in Lines 41–62. If the data are in columns other than those provided here, the column numbers will need to be modified. |
| Line 39 | Defines missing data as being coded '9'. If missing data have been coded as some other character, this line will need to be modified. |
| Lines 41–62 | These are sample dummy data. Once the program has been successfully run and the results below obtained, these lines may be either omitted or replaced with 'live' data. |
| Lines 64–102 | Impute missing values. |
| Lines 104–118 | Set up transformed AQoL variables. This procedure keeps the original data intact, thereby enabling the original data to be used for other analyses (e.g. computing simple summated AQoL HRQoL-scores). |
| Lines 120–179 | Recodes the AQoL variable responses to their respective utility weights. These lines should not be modified: any modification will cause invalid computation of utilities. |
| Lines 181–185 | Compute the disutilities for each AQoL dimension. Again, these lines should not be modified under any circumstances. |

¹⁵ It can be directly downloaded from the CHPE website <http://ariel.unimelb.edu.au/chpe/>

-
- Lines 187–191 Compute the utilities for each AQoL dimension.

 - Lines 195–200 Compute the AQoL utility score. These lines should not be modified under any circumstances.

 - Lines 202–204 Simply print out the AQoL dimension disutility utility scores and the AQoL utility scores for each case. These lines have been included to enable checking of the program with the small database in Lines 41–62. Once the program has been successfully run these two lines can be omitted.

Obtained results

Once the program has been typed in and saved (preferably as an *.sps file, but any extension can be used) it should be run using the provided dataset to check for errors. Although this dataset does not check all program lines, it enables a researcher to proceed with some confidence. The values were assigned as a test pattern and as such they have no particular meaning.

The obtained results should be identical to those shown in Figure 15. Note:

- Case #1 has no loss of utility. This is the highest value available for the AQoL. Any scores higher than 1.0 indicate an error in the data or the program.

- Cases #16–20 are reporting that death would be preferable to their HRQoL situation. Case #20 is the worst utility measured by the AQoL. Any scores below –0.04 will indicate an error in the data or program.

- The obtained DU...n values should be identical with those provided in the five lookup tables in Appendix II. This can always be used as an internal check when running the program.

Figure 15 Sample AQoL utility data analysis

ID	DU1	DU2	DU3	DU4	DU5	AQOL
001	.00	.00	.00	.00	.00	1.00
002	.21	.00	.00	.00	.00	1.00
003	.21	.00	.00	.12	.00	.88
004	.00	.10	.10	.12	.00	.73
005	.21	.00	.12	.06	.05	.80
006	.29	.21	.00	.18	.11	.60
007	.29	.21	.00	.30	.11	.52
008	.36	.10	.26	.06	.13	.57
009	.29	.36	.21	.14	.31	.32
010	.29	.38	.18	.37	.15	.29
011	.21	.38	.36	.31	.33	.20
012	.36	.31	.12	.39	.15	.34
013	.36	.21	.44	.32	.02	.33
014	.00	.74	.20	.27	.34	.12
015	.62	.29	.75	.23	.31	.11
016	.35	.36	1.00	.55	.31	.00
017	.75	.95	.82	.48	.02	-.01
018	.88	.92	.79	.56	.11	-.01
019	.77	.29	.51	.74	.92	-.03
020	1.00	1.00	1.00	1.00	1.00	-.04

3.3.2 Using the AQoL to estimate QALYs

The term Quality-adjusted life years (QALYs) refers to utility values existing over time; the formula is simply:

$$QALY = U * t$$

where U is the utility value of a health state and t is the length of time in years spent in that health state. If a person obtained an AQoL utility of 0.75 and they were in that health state for 5 years, the QALY-value would be $0.75 * 5 = 3.75$.

In order to use utilities to produce QALYs, three conditions must be met: (a) the utility weights must be based on preferences, (b) these preferences must be anchored on a 0.00–1.00 scale where 0.00=death and 1.00=good health (note that this allows health states worse than death) and (c) the preferences must form an interval scale, have a ‘strong interval’ property (Richardson 1994). Due to its design and construction procedures the AQoL meets (a) and (b). Demonstrating that the AQoL meets requirement (c) is difficult and has yet to be proven. In this, however, the AQoL is not alone: no test of this ‘strong interval’ property has been devised and no utility instruments have been proven to meet this requirement; it is assumed that existing instruments have it because of the scaling procedures used during preference elicitation and the combination rules followed to produce the utilities.

QALYs are usually used as an outcome measure in evaluations. The gain in QALYs due to the treatment would be the difference in QALYs between the intervention group and a suitable comparator group. Where costs are collected, the cost-per-QALY can be calculated enabling direct comparison with other interventions which have cost-per-QALY values known. Normally the intervention with the lowest cost-per-QALY gain would be preferred on the grounds of gaining a better health outcome for the least monetary input. For a more detailed discussion of QALYs and their uses readers should consult Drummond et al (1998) Methods for the Economic Evaluation of Health Care Programmes.

4 AQoL: Questions and Answers

We are committed to making it easy for everyone to use the AQoL through supporting researchers in their efforts to evaluate health programs. There are a number of questions which researchers commonly ask. In this section we present the most common questions we have been asked and provide answers. The questions cover two broad areas:

- Issues around economic evaluation and the use of generic utility instruments, of which the AQoL is an example; and
- Questions clarifying the uses and limitations of the AQoL instrument.

4.1 AQoL and economic evaluation

1. Does the AQoL measure costs, benefits or both?

AQoL and other multi-attribute utility (MAU) instruments have nothing to do with costs. They help to measure the benefits of health programs by quantifying the

quality (utility) of different health states in such a way that the quality and length of life may be combined as quality adjusted life years (QALYs).

2. What is the role of a MAU-instrument in economic evaluation?

MAU instruments measure 'utility' which is an index of the strength of a person's preference for a health state. This is usually measured on a scale on which zero (0.00) represents death and unity (1.00) is good health. A MAU instrument measures the utility of a health state. When the utility index number is multiplied by the number of years in this health state we obtain the number of QALYs.

Note that 'quality' actually means 'utility' which equates with the strength of a person's preference.

3. How can the quality of life (utility) be measured?

Utility is usually measured one of four ways. Each of these allows people to express the strength of their preference for a health state relative to death and good health. The four techniques are: (a) magnitude estimation; (b) standard gamble; (c) time trade-off; and (d) person trade-off.

The AQoL employs the time-trade-off (TTO), in which a person indicates the proportion of a given number of remaining years of life (usually defined as 10 years by the interviewer) that they would be prepared to give up in order to avoid living in the health state being measured. For example, if a person with a life expectancy of ten years on a dialysis machine was prepared to give up two of these years — that is 20% — to be in good health, then their utility score would be 0.8 (i.e. $1.0 - 0.2$, where '1' represents good health).

4. How do the values (utilities) given to us by the AQoL help answer the question: 'Should we undertake project/program X?'

The values may be used several ways to assist with decision-making about programs.

- Firstly, they may be the basis for cost utility analysis. With this, the utilities obtained from the AQoL will be multiplied by the number of years spent in the health state. This gives the number of quality adjusted life years (QALYs). The additional QALYs arising from a health intervention — the health program — are compared with the program costs to give the cost per QALY for the program. This may be compared with the cost per QALY of other programs. Unless there are other relevant factors — such as social equity — we would normally prefer programs where the cost per QALY is lowest as we may thereby obtain the largest number of QALYs – the best outcome – from a given budget. Importantly, cost utility analysis of this type cannot answer the question 'Should we undertake Program X?' as it only enables programs to be ranked by their cost per QALY. Of course a judgement may be made that the cost of a QALY is clearly too high or so low that a decision about the program is self-evident.

-
- Secondly, the utility values from the AQoL may be compared before and after a program. Where this is done for several programs, the different utilities can be compared. If one program is superior with respect to all possible criteria — cost, length of life, quality of life, and any other relevant factor — then it dominates the 'alternatives' and it should be preferred.
 - Thirdly, the AQoL can be used in program evaluation to produce a profile of health-related quality of life (HRQoL) as defined by the five different dimensions of health contained within it. Where measurement is made repeatedly, changes in health profiles can be tracked over time. This can be done for each dimension separately, or using the overall AQoL utility values.

5. Does the use of AQoL replace the need for a clinical trial or evaluation?

No; ideally economic evaluation incorporating utility measurement should be conducted alongside clinical trials. Economics is concerned with evaluating outcomes that are obtained from random control trials and other forms of experiments, quasi-experiments and pre-experimental research. It complements and does not replace other evaluative techniques. The result of an economic evaluation can be no better than the rigor of the clinical/epidemiological evaluation upon which it is based.

6. Can the non-utility values from the AQoL be stand-alone units of outcome?

Yes. This would occur where participants in a study completed the AQoL before and after an intervention. The difference between the obtained scores would provide a measure of the effect size of the program.

We expect most researchers will prefer to use the utility values. For cost utility analysis, any improvement would need to be multiplied by the length of time (e.g. the number of years) for which the improvement persisted. These length of time health gains represent the natural units of outcome from the intervention. In the case where an intervention had no impact upon the length of life and only upon the HRQoL then it would be appropriate to measure just the health state improvement using the AQoL.

7. Can the use of QALYs be justified?

Yes. The reason for using QALYs is that there is no alternative approach to combining the length and quality of life. Both of these are potential outcomes from health programs (i.e. programs may increase the length of life and the health-related quality of life). To assess the complete value (i.e. the real or true value) of an intervention these two outcomes should be combined and, implicitly or explicitly, used when decisions are eventually made to adopt or not to adopt a particular health program.

There are, however, a number of problems associated with the use of QALYs. Some of these are conceptual and arise from the fact that the quality and length of life are not the only outcomes from a program. For example, issues of process and the distribution of outcomes are important. Further there are a set of issues

concerning the accuracy of utility measurement. For these reasons, it is important that the role, strengths and weaknesses of QALYs be understood when evaluation results are interpreted. Despite these difficulties, the majority of health service researchers appear to accept that the measurement of QALYs represents an important development since they explicitly recognize the importance of health related quality of life and enable its measurement during program evaluation. The assumptions employed in measuring QALYs are transparent and may be subject to sensitivity analysis.

8. Isn't the composite approach to the measurement of QALYs superior to the use of a generic MAU- instrument such as the AQoL?

There are two ways of developing health state scenarios. With the composite approach, health state scenarios or vignettes — or complex health state descriptions — have been constructed and numerical values placed upon them using direct scaling. Under the multi-attribute utility (MAU) approach, health states are decomposed into a generic 'descriptive system' and a set of scale values corresponding with each possible health state in the descriptive system developed. As with the scenario-based approach, the values are obtained using one of the standard scaling techniques, *viz.* magnitude estimation, time trade-off, standard gamble, or, most recently, the person trade-off.

Both the approaches have strengths and weaknesses. The composite approach may include more context specific information and may describe a changing health scenario. It may include the risk and prognosis facing the patient. However, the validity of such scenarios is seldom (if ever) tested in the way in which the AQoL is being tested.¹⁶ In addition, because very few health states can be validated in this manner due to the time and cost of doing so, vignettes are limited in the range of health states covered and are insensitive. The use of vignettes in a series of studies increases the likelihood of comparability in the measurement of HRQoL. Generic instruments, on the other hand, are relatively cheap to administer and can be used across many health coalitions and for these reasons they can be completed by the same group of patients periodically during a longitudinal study in order to create a time profile of the HRQoL rather than relying upon a single point in time estimate.

4.2 Rationale for the AQoL

1. Are MAU instruments such as the AQoL necessary?

A large number of disease specific and a smaller number of generic QoL instruments exist (see, for example, Bowling 1991). These instruments do not weight the different dimensions of HRQoL by utility or the strength of people's preferences: different dimensions of HRQoL are simply added up to obtain an

¹⁶ The validity of the utilities obtained from the AQoL is being investigated through examining the exchange rate between the theoretical values and the obtained values on a life-death scale for a selected number of health state scenarios. To our knowledge this has not been previously investigated for any other utility instrument. For the results of this testing, reader should consult Richardson & Hawthorne (2000).

overall score.¹⁷ This implies there is a need for instruments which weight the different dimensions such that they can legitimately combined. Utility instruments achieve this weighting property through the elicitation of preferences for different health states, thus overcoming this criticism.

There are a number of situations in which it is necessary to know whether or not the overall quality of life has improved or deteriorated as the result of a health intervention; this is required for both summative program evaluation and for economic evaluation. With limited budgets we are commonly forced to select between programs. Consequently we must, implicitly or explicitly, compare the total benefits derived from competing programs. This requires an overall assessment of the program benefits; i.e. the derivation of a single index of HRQoL. Although, in principle, detailed utility studies could be carried out using the composite or vignette approach, in practice research budgets are limited and MAU instruments offer a low cost method for obtaining this information. Perhaps most importantly, MAU instruments have now gained world-wide acceptance and are being widely used. It is likely their use will continue to expand. It is therefore important to have instruments which minimise bias and maximise the likelihood of obtaining valid utility scores.

2 MAU instruments already exist. Do we need another?

In addition to the AQoL, there are nine other generic utility instruments in existence or which are/may be developed. These are the:

- Health Utility Index (HUI) Mark I, II, III (developed in Canada);
- Rosser-Kind index (UK);
- Quality of Wellbeing (QWB) instrument (USA);
- 15D (Finland);
- EuroQoL or EQ-5D (European, most commonly used with English weights);
- SF36 utility adaptation by Brazier, referred to as the SF6D (American/British);
- World Health Organization/World Bank DALY; and the
- World Health Organization's WHOQoL, which may eventually receive utility weights.

None of these instruments have been calibrated using Australian preference scores, although the available evidence shows that international variation in preferences is arguably relatively small. Several of the instruments are seriously compromised by the simplicity of their descriptive systems (HUI-I; Rosser-Kind; QWB; EuroQoL). Others have adopted a disease or impairment based

¹⁷ These comments do not apply to instruments developed using factor analysis to identify the various instrument scales.

descriptive system (HUI-II, III; 15D). Except for the HUI instruments the model employed by each of these instruments is problematical. The 15D and SF6D do not have satisfactory utility scores. None of the instruments have been subject to satisfactory validation against comparable instruments, and none has explicitly examined the relationship between HRQoL and the length of life. For a direct comparison of the psychometric and utility properties of the AQoL, and, HUI-III, EQ5D and 15D readers are referred to Hawthorne et al (2000), CHPE, Working Paper 102.

In sum, the construction of the AQoL was motivated by deficiencies in existing instruments.

3. What are the unique features of the AQoL?

- i. The AQoL is the only utility instrument the descriptive system of which was constructed using state-of-the-art psychometric instrument construction techniques.
- ii. The HUI instruments and the AQoL are the only utility instruments using a flexible multiplicative model for combining HRQoL dimensions.
- iii. The AQoL is the only instrument which independently models all the sub-dimensions of health and then combines these sub-models.¹⁸ By this device AQoL has maintained structural independence between health dimensions while simultaneously obtaining a high degree of descriptive sensitivity.
- iv. The AQoL project has undertaken a more exhaustive analysis of the exchange rate between HRQoL and life years than any other reported in the literature. At the time of writing, this enquiry is ongoing.
- v. Similarly, the AQoL is undergoing a large scale validation study by comparing its predicted values with the values obtained from five other generic instruments and from direct TTO self-assessment. Again, no similar study has been reported in the literature for any other instrument. A preliminary report on this study has been issued, and can be found at the CHPE website (Hawthorne et al 2000).

4.3 The AQoL and other QoL instruments

1. What is the relationship between the AQoL and non-utility instruments?

In addition to the utility instruments there are a very large number of disease specific and a smaller number of generic non-utility instruments (the SF36, the Nottingham Health Profile and the Sickness Impact Profile being examples of the latter). Most of these purport to measure health status. We recommend that researchers should incorporate all three levels of measurement. Each level provides different information about the effectiveness of an intervention and the

¹⁸ The HUI-III and SF6D model some sub-dimensions and combine with direct substitution values for some items.

different levels complement each other. Thus evaluation studies should include a disease specific instrument, one of the generic health-status instruments and a utility instrument.

The defining difference between the generic disease specific/health status instruments and the utility instruments is that the latter apply utility weights to different dimensions of health; these utilities (or dis-utilities) are then used in either a summative or multiplicative model to obtain a single index of HRQoL (or, more accurately, an index of the strength of a person's preference for this health state compared with full health and death). Without the utility weights the descriptive system of the AQoL could be (and has been (Lewis et al 1997)) used as a generic multi-attribute (psychometric) instrument where an overall score is obtained by summing the unweighted patient responses.

For use as a generic utility instrument the descriptive system must have certain important characteristics; *viz.* response categories for each item must be hierarchical (as in a Guttman scale); broad health dimensions must be orthogonal (there must be no double counting of health attributes); and there should be preference independence between dimensions (the preference score for one item or dimension must not depend upon the level of health defined by another dimension in the instrument; for example, preference dependence would occur if the dis-utility of pain increased when a person's social relationships were poor).

2. What is the relationship between the AQoL and other MAU instruments?

In principle the AQoL is similar to other MAU instruments; they all purport to measure the strength of preference for different health states on a 0 – 1 scale. In practice each of the existing MAU instruments differs in important respects.

Some conceptualise health in terms of disease characteristics: impairment and disability (HUI-I, II and III; 15D; DALY). Others have a heavier emphasis upon handicap: illness induced, or lack of capacity to carry out normal social activities (the AQoL, WHOQoL; SF36 and EuroQoL).

Even when broadly conceptualised in the same way, the descriptive systems of different generic instruments vary considerably with respect to the detail with which they describe different health dimensions. Instruments also differ with respect to the scaling (utility scoring) system adopted. Some are based upon the use of rating scales (15D and QWB); others have used the time-trade-off (the AQoL, EuroQoL and HUI instruments); one has used the person trade-off (the DALY); and one has used magnitude estimation (Rosser-Kind).

3. When should AQoL be used and when should another generic instrument be preferred?

The principal criterion for instrument selection should be whether or not the instrument is capable of measuring the health state(s) under investigation and is sufficiently sensitive to detect the changes that are likely to occur as a result of the intervention.

4.4 Scope and coverage of the AQoL

1. What does the AQoL measure?

As with any generic instrument, the AQoL measures the strength of a person's preference for a particular health state relative to the utility (preference) of good health and death, where these two 'end points' are assigned a value of unity ('1') and zero ('0') respectively.

Comparison of the AQoL with other instruments suggests that it is particularly sensitive in detecting the importance of social handicap. It also measures surrogate illness indicators (use of medication, and medical treatment); independent living indicators; physical and psychological domains. Like any QALY measure, the AQoL values do not vary with the context and process of health delivery (see 2 below). This is an important caveat. The repeated application of the AQoL can produce a profile through time, for example, a base line score can be obtained at the commencement of a stress management program. The re-application of the instrument at, say, two weekly intervals can track the impact of the program on the different dimensions of health through time as well as obtaining a final steady state score at the end of the program.

2. What does the AQoL omit that is relevant to an assessment of a health program?

Generic utility instruments do not take account of contextual factors, such as the medical or social environment of the patient. They do not measure the importance of prognosis or of anxiety associated with risk. The authors are unaware of any instrument which captures these potentially important aspects of a health program. The long term AQoL research agenda includes the investigation and inclusion of the severity of the initial health state and the impact of prognosis.

3. Is the AQoL intended to measure individual utility?

No. Economic evaluation — like all summative program evaluation — measures population outcome. That is, the extent to which a program 'works'. Similar to other outcome indicators, the utility scores are the average utilities obtained from a group of patients, trial participants and/or controls. Typically, there is a very large dispersion of individual results around the average score and, consequently, the average QALY value would be a highly unreliable predictor of the utility for any one individual person. It follows from this that QALYs are only appropriately used for assessing the overall impact of a program and not the benefit obtained by any given individual from that program. The purpose of cost utility is to rank the overall program and not to provide advice to clinicians or program managers about individuals.

4. Is the AQoL designed to measure individual utility or the utility of groups?

See #3 above. QALYs are designed to measure the average utility of a group of patients or program participants.

5. Can the AQoL measure changes through time attributable to normal disease progression?

Yes. This is one of the strengths of a simple generic instrument. It may be applied weekly, monthly or at any appropriate time interval.

6. Can the QALY values obtained from the AQoL be converted into dollars?

QALYs are intended to measure benefits, not costs. Typically cost utility analysis ranks programs by comparing the dollar cost per QALY obtained in the different programs of interest. In principle, a dollar value could be attached to each QALY and dollar costs compared with dollar benefits (thereby converting 'cost utility' analysis into 'cost benefits' analysis). The AQoL does not seek to do this. Cost effectiveness and cost utility analysis were introduced specifically to avoid the need to place a dollar value upon life *per se*. Some economists do convert life years or lives into dollar benefits using either the 'human capital' or 'willingness to pay', generally for risk reduction. These techniques are problematic.

7. Does AQoL measure costs?

No. The AQoL assists in the measurement of benefits which are then compared with costs in order to make a decisions.

4.5 Validation and psychometric properties

1. What is validation?

Validation is an ongoing process of testing an instrument in different ways and in different contexts to determine *whether or not the instrument measures what it purports to measure*. Unfortunately, the term 'validation' is commonly misunderstood; many people think it describes whether an instrument is or is not 'valid'. Instruments may be 'valid' in some circumstances but not in others. There are three kinds of evidence suggesting validity.

Content validity

Is defined as how well an instrument's items may be considered to be a representative sample of the universe which the researcher is trying to measure. In HRQoL measurement this might be the extent to which the items in an instrument cover the full domain of HRQoL; that is, whether or not the instrument includes items which enquire about each of the dimensions of health that are included in the underlying concept of HRQoL. Where content validity is determined by 'looking at the instrument' this is described as 'face validity'; although popular, apparent face validity does not confer content validity.

Construct validity

Construct validity indicates whether or not instrument items truly represent the underlying construct that is of interest. I.e. scales scores can be used to infer certain concepts. This implies the researcher has to define an underlying construct of whatever it is that is being measured. If no adequate construct is defined, the content of the instrument defines the construct that is being measured. If the construct is correctly represented by the instrument it is possible to draw a succession of inferences from the instrument concerning instrument scores in different contexts. Construct validity therefore involves an ongoing process of testing the instrument in different contexts.

Criterion validity

This describes the relationship between an instrument's scores and either other independent measures (the *criteria*) or other specific measures (*predictors*), where the criteria or predictors are the 'gold standard' for the measurement (e.g. in the case of breast cancer, the gold standard is the histopathological confirmation of cancer). In

the absence of a gold standard, confidence in criterion validity increases if the instrument has high correlation with each of the accepted *ex stant* instruments.

2. Which aspects of the AQoL need validation?

Both the descriptive system and the utility values associated with health states required validation. More specifically the descriptive system should be shown to have content, construct and criterion validity. Utility scores should correctly reflect the strength of people's preference for health states. As well, the model employed to combine the utility scores from different dimensions should also be valid.

3. Has the AQoL been validated?

The AQoL has apparent face validity for use in a wide range of contexts. As described elsewhere, AQoL was specifically created to ensure construct and content validity (Hawthorne et al 1996; Hawthorne et al 1997). At the time of writing criterion validity is being assessed by a comparison between AQoL scores and those obtained from the HUI-III; EQ5D; WHOQoL-Bréf, 15D and SF6D. A preliminary report on the findings (Hawthorne et al (2000)) can be obtained from the CHPE's webpage (<http://ariel.unimelb.edu.au/chpe/>)

4. How reliable is the AQoL

Reliability refers to the stability of measurement, either over time or between two different instruments measuring the same thing. Generally, there are three ways of assessing reliability: through internal consistency, test-retest and equivalence. Internal consistency is where the internal relationships between the different items are examined (usually reported as item-rest-of-test correlations, and as Cronbach α); test-retest involves administering the same instrument on two different occasions (about a week or fortnight apart; usually reported as Pearson r) and equivalence is where two different versions of the same instrument are administered and the relationship between scores examined (usually reported as Pearson r).

Both preliminary internal consistency and test-retest measures have been examined for the AQoL. For internal consistency, based on an additive model, the AQoL reliability was $\alpha = 0.87$, while for test-retest it was $r = 0.80$.

5. Have AQoL utility scores been validated?

This involves two separate issues.

- the validity of the TTO utility measurement technique; and
- the validity of AQoL scores after dimension utilities have been combined using a multiplicative model.

Validating scaling techniques is problematical as it is not possible to observe actual trade-offs between the quality and length of life which correspond with the trade-offs measured by the various scaling instruments. Consequently, validity has been determined primarily by face validity. Some have argued that the standard gamble should be regarded as the gold standard for utility measurement as its use assumes the axioms of von Neuman & Morgenstern; this appears to make the standard gamble results consistent with mainstream economic theory. However, as the axioms have been shown to be empirically incorrect we have not adopted this procedure. Rather, for the reasons outlined by Richardson (1994) and Dolan et al (1996), we have accepted the time trade off as having the greatest prima facie validity.

6. Why is modelling necessary: have utility models been validated?

The AQoL measures millions of health states. These cannot be measured individually and, like all other MAU instruments (except for the Rosser-Kind index) we have used a model to estimate utility scores from a limited number of observations. To date, most instruments have adopted an additive model in which the dis-utility associated with each response from each item is independently measured, and the overall dis-utility estimated or modelled as a weighted average of these dis-utilities, where the weights are also obtained empirically during the scaling survey. This additive model is probably invalid¹⁹ and the multiplicative model employed by the HUI and AQoL instruments is superior (Richardson & Hawthorne 1998). However there is no certainty that even this more flexible model does not introduce significant estimation bias. Like the HUI, the AQoL has tested its validity against a limited number of multi-attribute health states measured during the scaling survey. At the time of writing further validation is being obtained by the direct TTO measurement of patient scores. This will be used to modify the multiplicative model. This hybrid two stage AQoL modelling is the most flexible model used in any utility instrument scaling to date.

4.6 QoL and health promotion

1. Can AQoL Mark I be used for measuring the impact of health promotion upon population health?

Yes. For example, a road safety awareness program may prevent injury and death. Since the dis-utility associated with most injuries can be measured by the AQoL, any reduction in these injuries can be quantified. Similarly it is likely that the AQoL can measure the dis-utility associated with illnesses which would occur in the absence of an immunisation program. However, the AQoL is relatively insensitive to changes in health in the vicinity of full health. Our research indicates significant and important differences between 'normal', 'good', 'full' and 'best imaginable' health.

The AQoL is also a generic instrument designed for measuring the HRQoL of a representative cross-section of the Australian population. It is likely to be relatively insensitive to the particular health problems of specific sub-groups and particularly those which may be targeted by health promotion programs.

In general, the same rule applies as with any health program; viz. that prior to instrument selection any putative evaluation of an instrument's descriptive system should be examined to determine if it does or does not describe the expected health state before and after the health program.

2. Is a separate model needed to measure the outcome of health promotion?

See above. This depends upon the outcome of the program and the sub-population which receives the benefits of the program.

3. If a separate instrument is used for measuring the impact of health promotion will the utility scores be comparable with those obtained from other instruments?

No. Each utility instrument has different items, different dimensions of HRQoL, and different scaling properties. Given these differences there is no reason different instruments will necessarily provide comparable utilities. Empirical evidence on this

¹⁹ The source of invalidity is that an additive model prevents any health dimension causing significant disutility by itself. For example, suppose an instrument has five dimensions; under the additive model the maximum loss of utility (disutility) attributable to any one dimension would be 0.20 (1/5th of 1.00). A person who was, say, extremely depressed and suicidal could not achieve a utility of less than 0.80.

point is presented in Hawthorne et al (2000) where it is shown that very different estimates are obtained on the different utility instruments and that this is a function of the coverage of the instruments.

Although the different instruments all purport to measure the same underlying concept — the strength of preference for a health state — these health states will differ from instrument to instrument, given the differences referred to. At best, it could be expected there will be a moderate correlation ($r = 0.40$ — 0.60) between different utility values elicited from different instruments.

4. Is it possible for a MAU instrument such as the AQoL to measure all of the outcomes of health promotion?

There is still academic debate about the definition of health promotion and the nature of health promotional outcomes. The AQoL will be able to measure some but not all relevant outcomes; and these measures are confined to health improvements in the range of good health. We are currently researching this issue and will be constructing a modified AQoL (called AQoL II). AQoL II will be designed to specifically measure the impact of health programs on particular sub-populations and specifically the aged, adolescents, and NESB populations. In addition, during construction measurement of a broader range of issues relevant to health promotion will be considered, including empowerment and spirituality.

5. Could AQoL II be used to measure the same health states as AQoL I?

Generally, yes. It is an empirical question whether or not one instrument is more or less sensitive than the other in the context of any particular health state.

6. Could AQoL I and AQoL II produce different utility scores and different project ranking?

Yes, see #3 above. This would be true for any two HRQoL utility instruments. In principle utility scores should be moderately correlated. If they were not, this would imply the need for further research. Before the results of such research were known, both instruments should be used. If both produced similar rankings, confidence would be increased. If there was low correlation, confidence in the final outcome of the evaluation would decrease.

7. Does the construction of AQoL II require the full replication of the steps involved in the construction of AQoL I

Yes. The addition, deletion or supplementation of items could change the relative importance of all other items.

4.7 Ethical issues

1. General comment

The issues raised below are not specific to the AQoL. They concern the use of all generic MAU instruments and, more generally, the use of any HRQoL generic or disease specific instrument and the use of any utility scoring.

2. Can quality of life be reduced to a number?

There are different dimensions of health and combining them is like combining apples and oranges.

The index number produced by utility instruments is an index of *the strength of preference for different health states*. Combining dimensions is therefore analogous to combining the *preference* for apples with the *preference* for oranges; combining this is a valid procedure.

3. Can the quality and length of life be combined?

See #2 above.

4. Can people's utility scores – preferences – really be reduced to mathematical equations such as those employed by the AQoL?

Yes. The equations are nothing more than a sophisticated form of averaging individual utilities. The simplest form of averaging would be to add up the number of 'yes' answers in a simple questionnaire with yes/no answers to health related questions and then using the 'mathematical' equation which divides the total number of yeses by the number of people surveyed.

With a more sophisticated approach answers would be multiplied by an 'importance weight' and the 'mathematical equation' would be equivalent to the formula for a weighted average. This is the approach adopted by simple additive MAU models. Thirdly, the model may combine importance scores multiplicatively, ie scores may be multiplied together after adjusting for their importance and then the result scaled to constrain the answer to the 0-1 range. This is what the AQoL and HUI instruments do. The AQoL uses a two stage procedure in which the multiplicative model is used to combine items within each of the four dimensions used and then an overarching equation is used to obtain the final score.

It remains true, however, that the multiplicative model does impose a particular structure upon utility values and we are investigating the validity of this structure with the intention of adjusting predicted scores if necessary to more closely reflect the true pattern of utilities.

5. Doesn't the use of QALYs oversimplify measurement and ignore a large number of context specific and process related factors?

Yes. QALYs measure only two dimensions of outcome; viz. quality and length of life. However, these two dimensions are generally considered to be very important. This does not deny the importance of other process context and equity issues. Some of these are currently under investigation and are on the agenda for further reform of the AQoL. At present these factors must remain as 'intangibles' and their importance included subjectively in an overall assessment of a health program.

6. Who should judge the utility scores?

There has been vigorous debate over this issue in the literature. The AQoL currently incorporates the utility values of a representative cross-section of the Australian population. In this respect the AQoL has adopted the common practice of using community values.

Current research is underway to determine the correspondence between these and the self-assessments of hospital patients.

7. Can the average utility derived by the AQoL truly represent the utility of each individual?

No. Where individual choice is possible it is clearly better to allow individuals to select their own programs and to state their own preferences. Cost utility analysis is primarily useful for evaluating programs based on the 'common' or 'average' experience. For example, a new technology must either be installed or not installed in a hospital; a new procedure must be either included or not included in the Medicare schedule or in the benefits provided by a private health fund. In these cases it is necessary to make a collective choice and there is no mechanism for making such choices, other than consensus, which can reflect every person's preference. The ethical *strength* of the QALY procedure is that the final decision will be based upon

public preferences and the strength of these preferences and not upon a process which dis-enfranchises those who are affected by the decision.

8. Does CUA discriminate against the disabled by assigning them a lower utility score?

No. This argument was effectively used to discredit the prioritisation process adopted in the State of Oregon in the USA. The argument is, however, almost certainly wrong. Benefits in the CUA procedure are not derived from the utility of a person's long term health state before the procedure (as implied by the above argument) but from the *improvement* that results from an intervention. The improvement will generally be the same for the long term disabled and the remainder of the population. Further, when the total range of possible benefits is considered, the potential for health improvement by the disabled is greater and the potential QALY gain is larger for disabled persons since it may be possible to cure some disabilities.

In theory, there is a situation in which discrimination may occur, known as the 'double jeopardy' argument. This maintains that saving the life of a long term disabled citizen, for example a quadriplegic, will result in a smaller QALY gain than life saving for a person who is not disabled because the final QALY score for the disabled person will be lower. This theoretical argument has never been investigated nor has it ever been supported by those constructing QALY instruments. There is an intellectual challenge to reconcile the double jeopardy scenario with the *prima facie* (but not necessary) need for mathematical consistency in the manipulation of QALY scores and this challenge is being investigated at our Centre. The intellectual exercise, however, has no possible practical ramifications as long as the community rejects discrimination against the disabled. CUA must reflect and not impose community values.

References

- Anastasi A 1986, Evolving concepts of test validation, *Annual Review of Psychology*, 37: 1-15.
- Anastasi A 1976, *Psychological testing*, 4th ed, Macmillan Publishing, New York.
- Bagozzi R & Heatherton T 1994, A general approach to representing multifaceted personality constructs: application to state self-esteem, *Structural Equation Modelling*, 1(1): 35-67.
- Bowling A 1995, *Measuring disease: a review of disease-specific quality of life measurement scales*, Open University Press, Buckingham.
- Bowling A 1991, *Measuring health: a review of quality of life measurement scales*, Open University, Milton Keynes.
- Cohen J 1988, *Statistical power analysis for the behavioural sciences*, 2nd ed, Lawrence Erlbaum, Hillsdale.
- Cronbach J. & Meehl P 1955, 'Construct validity in psychological test', *Psychological Bulletin*, 52(4): 281-302.
- Dean A et al 1994, *EpilInfo version 6: a word processing, database and statistics program for epidemiology on microcomputers*, Centers for Disease Control and Prevention, Atlanta.
- Derogatis L 1992, *The affects balance scale: a preliminary guide to administration, scoring and application*, Clinical Psychometric Research Inc, Towson.
- Drummond M et al 1998, *Methods for the economic evaluation of health care programmes*, 2nd ed, Oxford University Press, Oxford.
- Dolan P & Kind P 1996, 'Inconsistency and health state valuations', *Soc Sci Med*, 42(4): 609-15.
- Dolan P et al 1994, *The measurement and valuation of health: first report on the main survey*, Centre for Health Economics, University of York, York.
- EuroQoLGroup 1990, EuroQoL - a new facility for the measurement of health-related quality of life, *Health Policy*, 16: 199-208.
- Fairclough D 1998, 'Methods of analysis for longitudinal studies of health-related quality of life', in *Quality of Life Assessment in Clinical Trials: Methods and Practice*, MJ Staquet, Hays RD & PM Fayers, Oxford University Press, Oxford.
- Feeny D, Furlong W & Torrance G 1996, *Health utilities index Mark 2 and Mark 3 (HUI2/3) 15-item questionnaire for self-administered, self-assessed usual health status*, Centre for Health Economics and Policy Analysis, McMaster University, Hamilton.
- Feeny D, Torrance G & Furlong W 1996, 'Health utilities index', in *Quality of life and pharmacoeconomics in clinical trials*, B. Spilker ed, Lippincott-Raven Publishers, Philadelphia.
- Gould A 1980, A new approach to the analysis of clinical drug trials with withdrawals, *Biometrics*, 36: 721-727.

-
- Guyatt GH et al 1996, 'Measurements in clinical trials: choosing the right approach', in *Quality of life and pharmacoeconomics in clinical trials*, B. Spilker, ed, Lippincott-Raven Publishers, Philadelphia, 41-48.
- Hawthorne G, Herrman H & Thomas R 2000, 'The validity of self-report health-related quality of life among those suffering mental illness', *Journal of Mental Health Policy and Economics*, 3, Supplement 1: S10.
- Hawthorne G et al 2000, *Life and death: theoretical and practical issues in utility measurement*, Centre for Health Program Evaluation, Working Paper 102, Melbourne.
- Hawthorne G et al 2000, *Construction and utility scaling of the assessment of quality of life (AQoL) instrument*, Centre for Health Program Evaluation, Working Paper 101, Melbourne.
- Hawthorne G, Richardson J & Osborne R 1999, 'The assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life, *Quality of Life Research*, 8: 209-224.
- Hawthorne G et al 1997, *The assessment of quality of life (AQoL) instrument: construction, initial validation and utility scaling*, Centre for Health Program Evaluation, Working Paper 66, Melbourne.
- Hawthorne G et al 1996, *The Australian multi-attribute utility (AMAU): construction and initial evaluation*, Centre for Health Program Evaluation, Working Paper 56, Melbourne.
- Hawthorne G & Richardson J 1995, 'An Australian MAU/QALY instrument: rationale and preliminary results' in *Proceedings Health Outcomes and Quality of Life Measurement Conference*, Canberra.
- Keeney R & Raiffa H 1976, *Decisions with multiple objectives: preferences and value trade-offs*, Wesley, New York.
- Kraemer H & Thieman S 1987, *How many subjects? Statistical power analysis in research*, Sage Publications, Newbury Park.
- Lennon R 1965, Assumptions underlying the use of content validity, *Educational & Psychological Measurement*, 16: 294-304.
- Lewis D et al 1997, 'Evaluation of the Cedar Court Rehabilitation Hospital health enhancement lifestyle program', in *2nd Annual Alternatives Therapies Symposium*, Orlando, Florida, USA.
- Mahoney F & Barthel D 1965, Functional evaluation: the Barthel index, *Maryland State Medical Journal*, 14: 61-65.
- McArdle J 1996, Current directions in structural factor analysis, *Current Directions in Psychological Science*, 5(1): 11-18.
- McCullagh P & Nelder J 1983, *Generalized linear models*, Chapman & Hall, London.
- Murray C & Lopez A 1996, *The global burden of disease*, World Health Organization, Geneva.
- Nordenfelt L 1994, ed *Concepts and measurement of quality of life in health care*, Kluwer Academic, Dordrecht.

-
- Orley J & Kuyken W 1994, *Quality of life assessment: international perspectives*, Springer-Verlag, Berlin.
- Pedhazur E & Schmelkin L 1991, *Measurement, design and analysis: an integrated approach*, Lawrence Erlbaum, Hillsdale.
- Raphael D 1996, 'Defining quality of life: eleven debates concerning its measurement', in *Quality of life in health promotion and rehabilitation*, R. Renwick, I. Brown, and M. Nagler eds, Sage Publications, Thousand Oaks.
- Richardson J 1994, 'Cost utility analysis: what should be measured?', *Social Science and Medicine*, 39(1): 7-21.
- Richardson J & Hawthorne G 1998, 'Difficulty with life and death: methodological issues and results from the utility scaling of the 'Assessment of Quality of Life' (AQoL) instrument', in *Economics and Health 1988*, The University of Sydney, School of Health Services Management, Sydney.
- Rummel R 1970, *Applied factor analysis*, Evanston, 24-27.
- Sintonen H 1995, *The 15D measure of health-related quality of life: feasibility, reliability and validity of its valuation system*, National Centre for Health Program Evaluation, Working Paper 42, Melbourne.
- Sintonen H.1994, *The 15D measure of health-related quality of life: reliability, validity and sensitivity of its health state descriptive system*, National Centre for Health Program Evaluation, Working Paper 41, Melbourne.
- Sneeuw K et al 1997a, 'Assessing quality of life after stroke: the value and limitations of proxy ratings', *Stroke*, 28: 1541-1549.
- Sneeuw K et al 1997b, Value of caregiver ratings in evaluating the quality of life of patients with cancer, *Journal of Clinical Oncology*, 15:1206-1217.
- Sprangers M & Aaronson, N 1992, The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review, *Journal of Clinical Epidemiology*, 45: 743-760.
- SPSS 1996, *Statistical package for the social sciences*, SPSS Inc, Chicago.
- Streiner D & Norman G 1995, *Health Measurement Scales: A practical guide to their development and use*, 2nd Edition, Oxford Medical Publications, Oxford.
- Torrance G 1986, 'Measurement of health state utilities for economic appraisal: a review', *Journal of Health Economics*, 5: 1-30.
- Torrance G et al.1996, Multiattribute utility function for a comprehensive health status classification system, Health Utilities Index Mark 2, *Medical Care*, 34(7) 702-22.
- Torrance G et al 1995, Multi-attribute preference functions: health utilities index, *Pharmacoeconomics*, 7(6): 503-520.
- Ware J et al 1993, *SF-36 health survey: manual and interpretation guide*. The Health Institute, New England Medical Centre, Boston.
- Winterfeldt D & Edwards W 1986, *Decision analysis and behavioural research*, Cambridge University Press, Cambridge.
- Wylie C & White B 1964, A measure of disability, *Archives of Environmental Health*, 8: 834-839.

Appendix 1 The AQoL instrument

Two versions of the AQoL instrument are provided here:

- Version A: This is for self-administration, e.g. in an interview or where the AQoL is posted to respondents.
- Version B: This is for telephone administration.

The differences between the two versions are related to the narrator: for self-administration the AQoL is presented in the first person, whereas for telephone administration it is presented in the third person.

Notes

1. Prior to administration of the AQoL, researchers should consult Section 2.3 for clarification of questions respondents may have.
2. For the telephone interview AQoL version greater detail is provided. This is to assist clear communication between the interviewer and the respondent by providing more guidance for each item.
 - All item stems include reference to the timeframe: 'in the last week'.
 - Optional illustrative examples are given using bullets and italics. Normally these cues should not be used. They should only be used where the respondent asks for clarification of an item. Incorporation of these into the telephone version was to try and standardize interviewers' responses to commonly asked questions.

The AQoL instrument: Version A²⁰

(For interview or mail self-administration.)

INSTRUCTIONS: ***This questionnaire has 15 questions and will take about ten minutes. The questions are about your health during the last week. Please circle the alternative that best describes you during the last week.***

- 1 Concerning my use of prescribed medicines:
- A. I do not or rarely use any medicines at all.
 - B. I use one or two medicinal drugs regularly.
 - C. I need to use three or four medicinal drugs regularly.
 - D. I use five or more medicinal drugs regularly.
- 2 To what extent do I rely on medicines or a medical aid? (NOT glasses or a hearing aid.)
(For example: walking frame, wheelchair, prosthesis etc.)
- A. I do not use any medicines and/or medical aids.
 - B. I occasionally use medicines and/or medical aids.
 - C. I regularly use medicines and/or medical aids.
 - D. I have to constantly take medicines or use a medical aid.
- 3 Do I need regular medical treatment from a doctor or other health professional?
- A. I do not need regular medical treatment.
 - B. Although I have some regular medical treatment, I am not dependent on this.
 - C. I am dependent on having regular medical treatment.
 - D. My life is dependent upon regular medical treatment.

²⁰ Copyright © Centre for Health Program Evaluation. All rights reserved. This material incorporating the AQoL instrument cannot be reproduced or applied without the prior written consent of the authors. This copyright is in the interests of all researchers as it ensures a standard instrument is available. This enables researchers to use the AQoL in full confidence they will be supported, and that their results will be directly comparable with those obtained by others.

-
- 4 Do I need any help looking after myself?
- A. I need no help at all.
 - B. Occasionally I need some help with personal care tasks.
 - C. I need help with the more difficult personal care tasks.
 - D. I need daily help with most or all personal care tasks.
- 5 When doing household tasks: *(For example, preparing food, gardening, using the video recorder, radio, telephone or washing the car)*
- A. I need no help at all.
 - B. Occasionally I need some help with household tasks.
 - C. I need help with the more difficult household tasks.
 - D. I need daily help with most or all household tasks.
- 6 Thinking about how easily I can get around my home and community:
- A. I get around my home and community by myself without any difficulty.
 - B. I find it difficult to get around my home and community by myself.
 - C. I cannot get around the community by myself, but I can get around my home with some difficulty.
 - D. I cannot get around either the community or my home by myself.
- 7 Because of my health, my relationships *(For example: with my friends, partner or parents)* generally:
- A. Are very close and warm.
 - B. Are sometimes close and warm.
 - C. Are seldom close and warm.
 - D. I have no close and warm relationships.
- 8 Thinking about my relationship with other people:
- A. I have plenty of friends, and am never lonely.
 - B. Although I have friends, I am occasionally lonely.
 - C. I have some friends, but am often lonely for company.
 - D. I am socially isolated and feel lonely.
- 9 Thinking about my health and my relationship with my family:
- A. My role in the family is unaffected by my health.
 - B. There are some parts of my family role I cannot carry out.
 - C. There are many parts of my family role I cannot carry out.
 - D. I cannot carry out any part of my family role.

-
- 10 Thinking about my vision, including when using my glasses or contact lenses if needed:
- A. I see normally.
 - B. I have some difficulty focusing on things, or I do not see them sharply.
For example: small print, a newspaper, or seeing objects in the distance.
 - C. I have a lot of difficulty seeing things. My vision is blurred.
For example: I can see just enough to get by with.
 - D. I only see general shapes, or am blind. *For example: I need a guide to move around.*
- 11 Thinking about my hearing, including using my hearing aid if needed:
- A. I hear normally.
 - B. I have some difficulty hearing or I do not hear clearly.
For example: I ask people to speak up, or turn up the TV or radio volume.
 - C. I have difficulty hearing things clearly. *For example: Often I do not understand what said. I usually do not take part in conversations because I cannot hear what is said.*
 - D. I hear very little indeed. *For example: I cannot fully understand loud voices speaking directly to me.*
- 12 When I communicate with others: *(For example: by talking, listening, writing or signing)*
- A. I have no trouble speaking to them or understanding what they are saying.
 - B. I have some difficulty being understood by people who do not know me. I have no trouble understanding what others are saying to me.
 - C. I am only understood by people who know me well. I have great trouble understanding what others are saying to me.
 - D. I cannot adequately communicate with others.
- 13 If I think about how I sleep:
- A. I am able to sleep without difficulty most of the time.
 - B. My sleep is interrupted some of the time, but I am usually able to go back to sleep without difficulty.
 - C. My sleep is interrupted most nights, but I am usually able to go back to sleep without difficulty.
 - D. I sleep in short bursts only. I am awake most of the night.
- 14 Thinking about how I generally feel:
- A. I do not feel anxious, worried or depressed.
 - B. I am slightly anxious, worried or depressed.
 - C. I feel moderately anxious, worried or depressed.
 - D. I am extremely anxious, worried or depressed.

15 How much pain or discomfort do I experience?

- A. None at all.
- B. I have moderate pain.
- C. I suffer from severe pain.
- D. I suffer unbearable pain.

The AQoL instrument: Version B²¹

(For telephone administration.)

INSTRUCTIONS:

Hello, my name is <name>, and I'm from the <organization>.

<Short introductory statement about the research project and its importance, confidentiality of material, right to stop interview at any time.>

This questionnaire has 15 questions and will take about ten minutes. The questions are about your health during the last week. Please listen carefully, and choose the answer that best describes you.

Question 1. Concerning your use of prescribed medicines in the last week.

- *[If the person asks 'What is a prescribed medicine', explain it refers to a medicine prescribed by a doctor, and it does not include over-the-counter drugs.]*

Would you say that:

1. You did not or rarely used any medicines at all.
2. You used one or two medicinal drugs regularly.
3. You needed to use three or four medicinal drugs regularly.
4. You used five or more medicinal drugs regularly.

²¹ Copyright © Centre for Health Program Evaluation. All rights reserved. This material incorporating the AQoL instrument cannot be reproduced or applied without the prior written consent of the authors. This copyright is in the interests of all researchers as it ensures a standard instrument is available. This enables researchers to use the AQoL in full confidence they will be supported, and that their results will be directly comparable with those obtained by others.

Question 2. To what extent did you rely on medicines or a medical aid in the last week, excluding your glasses or hearing aid.

- *[If the person asks for an example explain this refers to a walking frame, wheelchair, or prosthesis etc.]*
- *[If the person asks about 'medicines', explain this refers to all medicines used whether prescribed by a doctor, allied health professional or bought from a chemist.]*
- *[If the person asks about 'prosthesis', explain this refers to equipment used to replace a body part, such as an artificial arm.]*

Would you say that:

1. You did not use any medicines and/or medical aids.
2. You occasionally used medicines and/or medical aids.
3. You regularly used medicines and/or medical aids.
4. You had to constantly take medicines or use a medical aid.

Question 3. In the last week, did you need medical treatment from a doctor or other health professional?

Would you say:

1. You did not need regular medical treatment.
2. You had some regular medical treatment.
3. You were dependent on having regular medical treatment.
4. That your life was dependent on regular medical treatment.

Question 4. Did you need any help with personal care in the last week?

- *[If the person asks what is 'personal care', explain this refers to activities such as washing, dressing, personal grooming or going to the toilet.]*

Would you say:

1. You needed no help at all.
2. Occasionally you needed some help with personal care tasks.
3. You needed help with the more difficult personal care tasks.
4. You needed daily help with most or all personal care tasks.

Question 5. When doing household tasks during the last week, did you need any help?

- *[For example, with preparing food, gardening, using the video recorder, radio, telephone or washing the car.]*

Would you say:

1. You needed no help at all.
2. Occasionally you needed some help with household tasks.
3. You needed help with the more difficult household tasks.
4. You needed daily help with most or all household tasks.

Question 6. Thinking about how easily you got around your home and community in the last week.

Would you say:

1. You got around your home and community by yourself without any difficulty.
2. You found it difficult to get around your home and community by yourself.

-
3. You could not get around the community by yourself, but you got around your home with some difficulty.
 4. You could not get around either the community or your home by yourself.

Question 7. Were your personal relationships in the last week affected by your health.

- *[For example: with your partner or parents.]*

Would you say your relationships:

1. Were very close and warm.
2. Were sometimes close and warm.
3. Were seldom close and warm.
4. You had no close and warm relationships.

Question 8. Were your relationships with other people during the last week affected by your health.

Would you say:

1. That you had plenty of friends, and you were never lonely.
2. That although you have friends, you were occasionally lonely.
3. That you have some friends, but you were often lonely.
4. That you felt socially isolated and lonely.

Question 9. Thinking about your health and your relationship with your family in the last week.

Would you say:

1. Your role in the family was not affected by your health.
2. There were some parts of your family role you could not carry out.
3. There were many parts of your family role you could not carry out.
4. You could not carry out any part of your family role.

Question 10. Thinking about your vision in the last week.

- *[including when using your glasses or contact lenses if needed.]*

Would you say:

1. You saw normally.
2. You had some difficulty focusing on things, or you did not see them sharply.
[For example: small print, a newspaper, or seeing objects in the distance.]
3. You had a lot of difficulty seeing things and your vision was blurred.
[For example: you saw just enough to get by with.]
4. You only saw general shapes, or you are blind.
[For example: you needed a guide to move around.]

Question 11. Thinking about your hearing in the last week.

- *[Including using a hearing aid if needed]*

Would you say:

1. You heard normally.
2. You had some difficulty hearing, or did not hear clearly.
[For example: you asked people to speak up, or turn up the TV or radio volume.]
3. You had difficulty hearing things clearly.
[For example: Often you did not understand what was said. You usually did not take part in conversations because you could not hear what was said.]
4. You heard very little indeed.
[For example: you could not fully understand loud voices speaking directly to

you.]

-
- Question 12 When you communicated with others in the last week.
- *[For example: by talking, listening, writing or signing.]*
- Would you say:
1. You had no trouble speaking to others or understanding what they were saying.
 2. You had some difficulty being understood by people who did not know you. You had no trouble understanding what others were saying.
 3. You were only understood by people who knew you well. You had great trouble understanding what others were saying.
 4. You could not adequately communicate with others.
- Question 13. Thinking about how you slept in the last week. Would you say:
1. You slept without difficulty most of the time.
 2. Your sleep was interrupted some of the time, but you were usually able to go back to sleep without difficulty.
 3. Your sleep was interrupted most nights, but you were usually able to go back to sleep without difficulty.
 4. You slept in short bursts only. You were awake most of the night.
- Question 14. Thinking about how you generally felt in the last week. Would you say:
1. You did not feel anxious, worried or depressed.
 2. You were slightly anxious, worried or depressed.
 3. You felt moderately anxious, worried or depressed.
 4. You were extremely anxious, worried or depressed.
- Question 15. How much pain or discomfort did you experience in the last week. Would you say:
1. You had none at all.
 2. You had moderate pain.
 3. You suffered from severe pain.
 4. You suffered unbearable pain.

Thank you very much for answering these questions.

Appendix II

AQoL dimension disutility look-up tables

Table 1: Illness items, health state levels and disabilities

<i>Items & health state levels</i>				<i>Items & health state levels</i>			
<i>1</i>	<i>2</i>	<i>3</i>	<i>Disutilities</i> ¹	<i>1</i>	<i>2</i>	<i>3</i>	<i>Disutilities</i> ¹
1	1	1	0.0000	3	1	1	0.2083
1	1	2	0.0946	3	1	2	0.2860
1	1	3	0.2508	3	1	3	0.4142
1	1	4	0.5700	3	1	4	0.6763
1	2	1	0.1856	3	2	1	0.3607
1	2	2	0.2651	3	2	2	0.4260
1	2	3	0.3964	3	2	3	0.5338
1	2	4	0.6647	3	2	4	0.7541
1	3	1	0.3222	3	3	1	0.4728
1	3	2	0.3907	3	3	2	0.5290
1	3	3	0.5036	3	3	3	0.6218
1	3	4	0.7345	3	3	4	0.8113
1	4	1	0.6900	3	4	1	0.7748
1	4	2	0.7285	3	4	2	0.8065
1	4	3	0.7921	3	4	3	0.8587
1	4	4	0.9222	3	4	4	0.9654
2	1	1	0.1279	4	1	1	0.3900
2	1	2	0.2121	4	1	2	0.4529
2	1	3	0.3512	4	1	3	0.5568
2	1	4	0.6353	4	1	4	0.7690
2	2	1	0.2931	4	2	1	0.5134
2	2	2	0.3639	4	2	2	0.5663
2	2	3	0.4808	4	2	3	0.6536
2	2	4	0.7196	4	2	4	0.8320
2	3	1	0.4147	4	3	1	0.6043
2	3	2	0.4757	4	3	2	0.6498
2	3	3	0.5762	4	3	3	0.7249
2	3	4	0.7817	4	3	4	0.8784
2	4	1	0.7421	4	4	1	0.8488
2	4	2	0.7764	4	4	2	0.8745
2	4	3	0.8330	4	4	3	0.9168
2	4	4	0.9487	4	4	4	1.0032

1 = Based on means

Table 2: Independent living items, health state levels and disabilities

<i>Items & health state levels</i>				<i>Items & health state levels</i>			
<i>4</i>	<i>5</i>	<i>6</i>	<i>Disutilities</i> ¹	<i>4</i>	<i>5</i>	<i>6</i>	<i>Disutilities</i> ¹
1	1	1	0.0000	3	1	1	0.2700
1	1	2	0.2054	3	1	2	0.4249
1	1	3	0.2615	3	1	3	0.4672
1	1	4	0.6300	3	1	4	0.7452
1	2	1	0.1244	3	2	1	0.3639
1	2	2	0.3066	3	2	2	0.5012
1	2	3	0.3563	3	2	3	0.5388
1	2	4	0.6831	3	2	4	0.7853
1	3	1	0.1749	3	3	1	0.4020
1	3	2	0.3476	3	3	2	0.5322
1	3	3	0.3948	3	3	3	0.5678
1	3	4	0.7046	3	3	4	0.8015
1	4	1	0.5100	3	4	1	0.6547
1	4	2	0.6201	3	4	2	0.7377
1	4	3	0.6501	3	4	3	0.7604
1	4	4	0.8476	3	4	4	0.9094
2	1	1	0.1032	4	1	1	0.6700
2	1	2	0.2893	4	1	2	0.7502
2	1	3	0.3401	4	1	3	0.7720
2	1	4	0.6740	4	1	4	0.9159
2	2	1	0.2159	4	2	1	0.7186
2	2	2	0.3810	4	2	2	0.7897
2	2	3	0.4260	4	2	3	0.8091
2	2	4	0.7221	4	2	4	0.9366
2	3	1	0.2617	4	3	1	0.7383
2	3	2	0.4182	4	3	2	0.8057
2	3	3	0.4609	4	3	3	0.8241
2	3	4	0.7417	4	3	4	0.9450
2	4	1	0.5653	4	4	1	0.8691
2	4	2	0.6650	4	4	2	0.9120
2	4	3	0.6922	4	4	3	0.9237
2	4	4	0.8712	4	4	4	1.0008

¹ = Based on means

Table 3: Social relationships items, health state levels and disabilities

<i>Items & health state levels</i>				<i>Items & health state levels</i>			
<i>7</i>	<i>8</i>	<i>9</i>	<i>Disutilities</i> ¹	<i>7</i>	<i>8</i>	<i>9</i>	<i>Disutilities</i> ¹
1	1	1	0.0000	3	1	1	0.2891
1	1	2	0.1014	3	1	2	0.3623
1	1	3	0.2049	3	1	3	0.4370
1	1	4	0.6900	3	1	4	0.7872
1	2	1	0.0617	3	2	1	0.3337
1	2	2	0.1572	3	2	2	0.4025
1	2	3	0.2545	3	2	3	0.4728
1	2	4	0.7108	3	2	4	0.8022
1	3	1	0.1241	3	3	1	0.3787
1	3	2	0.2135	3	3	2	0.4423
1	3	3	0.3046	3	3	3	0.5090
1	3	4	0.7317	3	3	4	0.8173
1	4	1	0.6500	3	4	1	0.7583
1	4	2	0.6880	3	4	2	0.7858
1	4	3	0.7268	3	4	3	0.8138
1	4	4	0.9085	3	4	4	0.9450
2	1	1	0.1234	4	1	1	0.7300
2	1	2	0.2128	4	1	2	0.7602
2	1	3	0.3040	4	1	3	0.7910
2	1	4	0.7315	4	1	4	0.9354
2	2	1	0.1778	4	2	1	0.7484
2	2	2	0.2619	4	2	2	0.7768
2	2	3	0.3477	4	2	3	0.8058
2	2	4	0.7498	4	2	4	0.9416
2	3	1	0.2328	4	3	1	0.7670
2	3	2	0.3115	4	3	2	0.7936
2	3	3	0.3918	4	3	3	0.8207
2	3	4	0.7683	4	3	4	0.9479
2	4	1	0.6962	4	4	1	0.9235
2	4	2	0.7297	4	4	2	0.9348
2	4	3	0.7639	4	4	3	0.9464
2	4	4	0.9241	4	4	4	1.0005

1 = Based on means

Table 4: Physical senses items, health state levels and disabilities

<i>Items & health state levels</i>				<i>Items & health state levels</i>			
<i>10</i>	<i>11</i>	<i>12</i>	<i>Disutilities</i> ¹	<i>10</i>	<i>11</i>	<i>12</i>	<i>Disutilities</i> ¹
1	1	1	0.0000	3	1	1	0.1181
1	1	2	0.1226	3	1	2	0.2320
1	1	3	0.1921	3	1	3	0.2965
1	1	4	0.5600	3	1	4	0.6381
1	2	1	0.0860	3	2	1	0.1980
1	2	2	0.2023	3	2	2	0.3059
1	2	3	0.2681	3	2	3	0.3671
1	2	4	0.6169	3	2	4	0.6910
1	3	1	0.1625	3	3	1	0.2690
1	3	2	0.2731	3	3	2	0.3717
1	3	3	0.3357	3	3	3	0.4299
1	3	4	0.6675	3	3	4	0.7380
1	4	1	0.3400	3	4	1	0.4388
1	4	2	0.4375	3	4	2	0.5243
1	4	3	0.4926	3	4	3	0.5756
1	4	4	0.7850	3	4	4	0.8471
2	1	1	0.0595	4	1	1	0.4100
2	1	2	0.1777	4	1	2	0.5023
2	1	3	0.2446	4	1	3	0.5545
2	1	4	0.5993	4	1	4	0.8313
2	2	1	0.1424	4	2	1	0.4747
2	2	2	0.2545	4	2	2	0.5622
2	2	3	0.3179	4	2	3	0.6117
2	2	4	0.6542	4	2	4	0.8741
2	3	1	0.2161	4	3	1	0.5323
2	3	2	0.3228	4	3	2	0.6155
2	3	3	0.3831	4	3	3	0.6626
2	3	4	0.7031	4	3	4	0.9122
2	4	1	0.3872	4	4	1	0.6658
2	4	2	0.4812	4	4	2	0.7391
2	4	3	0.5344	4	4	3	0.7806
2	4	4	0.8163	4	4	4	1.0006

¹ = Based on means

Table 5: Psychological wellbeing items, health state levels and disabilities

<i>Items & health state levels</i>				<i>Items & health state levels</i>			
<i>13</i>	<i>14</i>	<i>15</i>	<i>Disutilities</i> ¹	<i>13</i>	<i>14</i>	<i>15</i>	<i>Disutilities</i> ¹
1	1	1	0.0000	3	1	1	0.0240
1	1	2	0.0853	3	1	2	0.1077
1	1	3	0.2558	3	1	3	0.2751
1	1	4	0.8200	3	1	4	0.8288
1	2	1	0.0465	3	2	1	0.0696
1	2	2	0.1287	3	2	2	0.1503
1	2	3	0.2932	3	2	3	0.3117
1	2	4	0.8370	3	2	4	0.8454
1	3	1	0.0657	3	3	1	0.0884
1	3	2	0.1466	3	3	2	0.1679
1	3	3	0.3085	3	3	3	0.3268
1	3	4	0.8440	3	3	4	0.8523
1	4	1	0.3300	3	4	1	0.3479
1	4	2	0.3935	3	4	2	0.4102
1	4	3	0.5205	3	4	3	0.5348
1	4	4	0.9406	3	4	4	0.9471
2	1	1	0.0235	4	1	1	0.2200
2	1	2	0.1073	4	1	2	0.2908
2	1	3	0.2747	4	1	3	0.4323
2	1	4	0.8286	4	1	4	0.9004
2	2	1	0.0692	4	2	1	0.2586
2	2	2	0.1499	4	2	2	0.3268
2	2	3	0.3114	4	2	3	0.4632
2	2	4	0.8453	4	2	4	0.9145
2	3	1	0.0880	4	3	1	0.2745
2	3	2	0.1675	4	3	2	0.3416
2	3	3	0.3264	4	3	3	0.4760
2	3	4	0.8522	4	3	4	0.9203
2	4	1	0.3475	4	4	1	0.4938
2	4	2	0.4099	4	4	2	0.5465
2	4	3	0.5345	4	4	3	0.6519
2	4	4	0.9470	4	4	4	1.0004

¹ = Based on means

Note:

Perspicuity rather than parsimony guided the development of the computer code in this section in an effort to make it universally accessible. Programmers who seek more elegance should consider RECODE and DO REPEAT statements. Where these are used they will considerably shorten the program.

- An electronic version of this program can be downloaded from the CHPE website (<http://ariel.unimelb.edu.au/chpe/>).

```

1 *****
2 *** REM This file analyses the AQoL and produces utilities for each dimension ***
3 *** and the instrument overall. ***
4 *** REM The AQoL uses D2, D3, D4 & D5 ***
5 *** when calculating the utilities, ***
6 *** although dimension utilities are calculated for D1. ***
7 *** REM The dimensions are scaled on a ***
8 *** "Dimension Worst Health State - Dimension Best Health State" scale ***
9 *** where DWHS = 0.00 and DBHS = 1.00. ***
10 *** These are not strict utility values as they have not been evaluated on ***
11 *** a life-death scale. ***
12 *** REM The AQoL utility scores are scaled such that the: ***
13 *** "AQoL worst health state" = -0.04 (i.e. this is worse than Death, where ***
14 *** Death = 0.00). ***
15 *** "AQoL best health state" = 1.00 (i.e. this is good health) ***
16 *** REM Copyright. Version 3. Release date: September 1999. ***
17 *** REM Version 3 is an interim release which replaces Versions 1 & 2. ***
18 *** This version will increase utility scores by about 3% ***
19 *** at the bottom end of the utility scale when compared with Version 1. ***
20 *** REM Note that missing data are handled by imputing values within each ***
21 *** dimension. Regression analysis should not be used because this may ***
22 *** violate the statistical independence of each dimension and result in ***
23 *** double-counting. ***
24 *** REM THIS IS AN INTERIM RELEASE WHICH IS SUBJECT TO REVISION WITHOUT NOTICE.
25 ***
26 *** REM RESEARCHERS SHOULD CHECK WITH THE AQOL TEAM FOR SUBSEQUENT
27 *** MODIFICATION ***
28 *** REM Note: "AQoL1" etc. are the variables in your questionnaire or database; ***
29 *** you will need to replace these with your variable names. ***
30 *** This only applies to the first 15 COMPUTE statements. ***
31 *** Once these have been changed, your variables are no longer used ***
32 *** anywhere in the program. ***
33 *****
34 Data list /ID 1-3
35 Q1 5 Q2 6 Q3 7
36 Q4 8 Q5 9 Q6 10
37 Q7 11 Q8 12 Q9 13
38 Q10 14 Q11 15 Q12 16
39 Q13 17 Q14 18 Q15 19.
40 Missing values All (9).
41 Begin data.
42 001 1111111111111111
43 002 1211111111111111
44 003 1211111111211111
45 004 112111121121112
46 005 121112112111211
47 006 211121112122121
48 007 211121113132121

```

```
49      008 222112222111222
50      009 211232122211333
51      010 212222213322221
52      011 122223123222332
53      012 221221312412222
54      013 221123322233113
55      014 114311131234321
56      015 332122243122233
57      016 131234444133232
58      017 244344332423112
59      018 344433421143123
60      019 412123333344344
61      020 444444444444444
62      End data.
63
64      Compute ILLmiss = Nmiss (Q1 to Q3).
65      Do if ILLmiss = 1.
66      Do repeat
67          A = Q1 to Q3.
68      If (Missing (A)) A = RND(Mean (Q1 to Q3)).
69      End repeat.
70      End if.
71
72      Compute ADLmiss = Nmiss (Q4 to Q6).
73      Do if ADLmiss < 2.
74      Do repeat
75          A = Q4 to Q6.
76      If (Missing (A)) A = RND(Mean (Q4 to Q6)).
77      End repeat.
78      End if.
79
80      Compute SOCmiss = Nmiss (Q7 to Q9).
81      Do if SOCmiss < 2.
82      Do repeat
83          A = Q7 to Q9.
84      If (Missing (A)) A = RND(Mean (Q7 to Q9)).
85      End repeat.
86      End if.
87
88      Compute PHYmiss = Nmiss (Q10 to Q12).
89      Do if PHYmiss < 2.
90      Do repeat
91          A = Q10 to Q12.
92      If (Missing (A)) A = RND(Mean (Q10 to Q12)).
93      End repeat.
94      End if.
95
96      Compute PSYmiss = Nmiss (Q13 to Q15).
97      Do if PSYmiss < 2.
98      Do repeat
```

99 A = Q13 to Q15.
100 If (Missing (A)) A = RND(Mean (Q13 to Q15)).
101 End repeat.
102 End if.
103
104 Compute U1 = Q1.
105 Compute U2 = Q2.
106 Compute U3 = Q3.
107 Compute U4 = Q4.
108 Compute U5 = Q5.
109 Compute U6 = Q6.
110 Compute U7 = Q7.
111 Compute U8 = Q8.
112 Compute U9 = Q9.
113 Compute U10 = Q10.
114 Compute U11 = Q11.
115 Compute U12 = Q12.
116 Compute U13 = Q13.
117 Compute U14 = Q14.
118 Compute U15 = Q15.
119
120 If (U1 eq 1)U1 = 0.000.
121 If (U1 eq 2)U1 = 0.328.
122 If (U1 eq 3)U1 = 0.534.
123 If (U1 eq 4)U1 = 1.000.
124 If (U2 eq 1)U2 = 0.000.
125 If (U2 eq 2)U2 = 0.269.
126 If (U2 eq 3)U2 = 0.467.
127 If (U2 eq 4)U2 = 1.000.
128 If (U3 eq 1)U3 = 0.000.
129 If (U3 eq 2)U3 = 0.166.
130 If (U3 eq 3)U3 = 0.440.
131 If (U3 eq 4)U3 = 1.000.
132 If (U4 eq 1)U4 = 0.000.
133 If (U4 eq 2)U4 = 0.154.
134 If (U4 eq 3)U4 = 0.403.
135 If (U4 eq 4)U4 = 1.000.
136 If (U5 eq 1)U5 = 0.000.
137 If (U5 eq 2)U5 = 0.244.
138 If (U5 eq 3)U5 = 0.343.
139 If (U5 eq 4)U5 = 1.000.
140 If (U6 eq 1)U6 = 0.000.
141 If (U6 eq 2)U6 = 0.326.
142 If (U6 eq 3)U6 = 0.415.
143 If (U6 eq 4)U6 = 1.000.
144 If (U7 eq 1)U7 = 0.000.
145 If (U7 eq 2)U7 = 0.169.
146 If (U7 eq 3)U7 = 0.396.
147 If (U7 eq 4)U7 = 1.000.
148 If (U8 eq 1)U8 = 0.000.

149 If (U8 eq 2)U8 = 0.095.
150 If (U8 eq 3)U8 = 0.191.
151 If (U8 eq 4)U8 = 1.000.
152 If (U9 eq 1)U9 = 0.000.
153 If (U9 eq 2)U9 = 0.147.
154 If (U9 eq 3)U9 = 0.297.
155 If (U9 eq 4)U9 = 1.000.
156 If (U10 eq 1)U10 = 0.000.
157 If (U10 eq 2)U10 = 0.145.
158 If (U10 eq 3)U10 = 0.288.
159 If (U10 eq 4)U10 = 1.000.
160 If (U11 eq 1)U11 = 0.000.
161 If (U11 eq 2)U11 = 0.253.
162 If (U11 eq 3)U11 = 0.478.
163 If (U11 eq 4)U11 = 1.000.
164 If (U12 eq 1)U12 = 0.000.
165 If (U12 eq 2)U12 = 0.219.
166 If (U12 eq 3)U12 = 0.343.
167 If (U12 eq 4)U12 = 1.000.
168 If (U13 eq 1)U13 = 0.000.
169 If (U13 eq 2)U13 = 0.107.
170 If (U13 eq 3)U13 = 0.109.
171 If (U13 eq 4)U13 = 1.000.
172 If (U14 eq 1)U14 = 0.000.
173 If (U14 eq 2)U14 = 0.141.
174 If (U14 eq 3)U14 = 0.199.
175 If (U14 eq 4)U14 = 1.000.
176 If (U15 eq 1)U15 = 0.000.
177 If (U15 eq 2)U15 = 0.104.
178 If (U15 eq 3)U15 = 0.312.
179 If (U15 eq 4)U15 = 1.000.
180
181 Compute DU1 = (1.1641*(1-(1-0.3350*U1)*(1-0.5927*U2)*(1-0.4896*U3))).
182 Compute DU2 = (1.0989*(1-(1-0.6097*U4)*(1-0.4641*U5)*(1-0.5733*U6))).
183 Compute DU3 = (1.0395*(1-(1-0.7023*U7)*(1-0.6253*U8)*(1-0.6638*U9))).
184 Compute DU4 = (1.6556*(1-(1-0.2476*U10)*(1-0.2054*U11)*(1-0.3382*U12))).
185 Compute DU5 = (1.2920*(1-(1-0.1703*U13)*(1-0.2554*U14)*(1-0.6347*U15))).
186
187 Compute UD1 = 1-DU1.
188 Compute UD2 = 1-DU2.
189 Compute UD3 = 1-DU3.
190 Compute UD4 = 1-DU4.
191 Compute UD5 = 1-DU5.
192
193 *** REM This model uses W = 1.04.
194
195 Compute AQOL = ((1.04*((1-(0.613*0))*
196 (1-(0.841*DU2))*
197 (1-(0.855*DU3))*
198 (1-(0.931*DU4))*

199 (1-(0.997*DU5))) - 0.04).
200 Execute.
201
202 List
203 /Variables = ID DU1 DU2 DU3 DU4 DU5 UAQOL.
204 Execute.