The Australian Economic Review, vol. 37, no. 1, pp. 62–88

Measurement of the Quality of Life for Economic Evaluation and the Assessment of Quality of Life (AQoL) Mark 2 Instrument

Jeff Richardson[†], Neil Atherton Day[‡], Stuart Peacock[†] and Angelo Iezzi[†] *
[†] Health Economics Unit, Monash University
[‡] Centre for Program Evaluation, The University of Melbourne

Abstract

Including the quality of life in the economic assessment of health and medical services is well established in the literature and a number of multi-attribute utility (MAU) instruments are available which purport to measure health state utilities. One of these, the Assessment of Quality of Life (AQoL) instrument was developed in Australia and uses Australian importance weights. The present article discusses some of the methodological problems encountered by existing instruments. It outlines the construction of the AQoL Mark 2 and the methodological innovations which have attempted to overcome some of these problems.

Technical and other details may be obtained in Richardson et al. (2003a, 2003b, 2003c) and Peacock et al. (2003). These papers may be accessed from the Health Economics Unit web site at <http://heu.buseco.monash.edu.au>.

* The authors would like to acknowledge the financial assistance of the Victorian Health Promotion Foundation and the Population Health Division of the Commonwealth Department of Health and Ageing. We also benefited greatly from the comments of two anonymous reviewers.

1. Introduction

The objective of this article is to introduce the Assessment of Quality of Life (AQoL) Mark 2 multi-attribute utility (MAU) instrument, to outline the challenges it faced and the methodological innovations which sought to meet these challenges. The article commences with a brief discussion of the purpose of MAU instruments and some of the reasons for concern about the current generation of instruments. The second section outlines the theory of instrument construction and the various innovations in the methodology for the construction of AQoL 2. Section 3 describes the modelling of the AQoL 2 descriptive system and its 'scaling'—calibration. Results from the two construction surveys and interviews are presented in Section 4 and the algorithm for estimating utility scores is reported. Likely extensions to the instrument are discussed briefly in Section 5.

1.1 MAU Instruments and Their Purpose

Before the development of cost utility analysis (CUA), the economic evaluation of health services either ignored quality of life (QoL) or treated QoL as an 'intangible' that could be noted and described but not precisely quantified. CUA has attempted to overcome this deficit by adopting the quality-adjusted life year (QALY) as the unit of output in cost-effectiveness studies. The approach may be described as 'quasi-utilitarian' as its most fundamental assumption is that a year of full health results in the same level of utility for everyone.

Published by Blackwell Publishing Asia Pty Ltd

The assumption is a sufficient condition for the summation and comparison of utilities. An alternative and more defensible interpretation of the assumption is that CUA makes the normative judgement that, all else equal, the utility a person receives from full health *should* be treated as being equal (interestingly, the ethical assumption made by Jeremy Bentham, the originator of utilitarianism).

In CUA 'best health' is assigned a value of 1.00 and death a value of 0.00. Consequently, (positive) utility varies between 0.00 and 1.00 and QALYs may be calculated by multiplying life years by the numerical value of the health state utility.

The measurement of 'utility' requires two tasks. First, the health state under investigation must be described. Second, a scaling technique which purports to measure utility must be used to attach numerical values to the health state. The two methods most widely accepted as measuring utility are the standard gamble (SG) and the time trade-off (TTO) techniques.[1] These two tasks—description and measurement—may themselves be carried out in one of two ways. First, in the 'holistic' or composite approach to measurement, the relevant health states are described in a series of vignettes, or scenarios. These are then rated using the selected scaling instrument to obtain a 'utility' index which is used to calculate QALYs. The construction of the health scenarios and the rating exercise both require surveys. Normally, patients who have experienced the health states are consulted for scenario construction and a random sample of the population is used for the weighting.

The second, 'decomposed', approach requires the preliminary construction of a generic MAU QoL instrument which is capable of describing numerous health states and assigning a utility score to each of these. The first stage in the construction of an MAU instrument is therefore the construction of the 'descriptive system'. This involves the decomposition of a particular concept of health and describing each of the resulting attributes (dimensions or constituent parts) of the concept using one or more 'items'; that is, by a series of questions, each with multiple responses, which describe the dimension and the intensity of the health state experienced. To convert the multi-attribute descriptive system into an MAU instrument, a scoring algorithm is created which can convert any combination of item responses into an index of utility. This is normally achieved by measuring a limited number of multi-attribute health states and using these to calibrate a model which is then used to infer the utility values of every other health state in the descriptive system.[2] The model may be derived either by econometric analysis of the observed utilities or by the use of decision analytic techniques to fit a simple additive or a multiplicative model.[3] The fully scaled MAU instrument may then be used to estimate the utility of health states.

Both approaches have strengths and weaknesses. Holistic measurement permits a description which is tailored to a particular health state. Unique aspects of the health state, its context, its consequences, the process of health care delivery, risk and prognosis may all be included in the vignette. Validation of health state-specific vignettes, however, is seldom, if ever, carried out. In contrast, the descriptive system of the MAU approach may be unable to capture many of the nuances of the health state and be incapable of capturing the importance of the process or context. However, this approach should, in principle, be based upon a descriptive system, the reliability and validity of which can be investigated using standard procedures. After construction, the use of an MAU instrument is inexpensive and easy and allows the rapid estimation of utilities in the context of a longitudinal trial. This means that it is feasible to construct a time profile of each of the dimensions of health included in the instrument. Because of these respective strengths and weaknesses both techniques have a role in CUA.

1.2 Problems with MAU Instruments

To date, only a handful of generic instruments have attempted to measure utility. (These are described and contrasted in Hawthorne, Richardson and Day 2001.) Each of these has particular strengths. However each has

limitations. These include an ad hoc approach to the construction of the descriptive system, the adoption of a limited concept of QoL (and, more specifically, the exclusion of important social elements), the use of a rating scale to obtain utility scores[4] and overly simplistic modelling. While there has been some limited discussion of construct validity, other MAU instruments have not demonstrated this property. More surprisingly, there has been little recognition in the economics literature of the need for rigorous validation studies—testing whether or not QALYs and MAU instruments measure what they purport to measure. There are now a large number of empirical studies which include both utility and disease-specific instruments (see Brazier et al. 1999 for a review). When the scores from the instrument of interest correlate with other instrument scores the assertion is generally made that the instrument has been 'validated'. At best, however, correlational evidence represents weak and context-specific 'validation'. It does not demonstrate the existence of a 'strong interval property' (Richardson 2002), namely the requirement that a 10 per cent increase in the numerical value of the 'utility' index is equivalent to a 10 per cent increase in life years or (otherwise equal) lives saved. Indeed, this property has been virtually ignored.

While there is no criterion test of the strong interval property, the plausibility of utility scores may be investigated by determining the implications of a utility score for the willingness to sacrifice life. The result of one such test of two instruments is reported in Table 1. Published values for the original McMaster (HUI 1) and the Quality of Wellbeing (QWB) instruments (column 1) were used to calculate the number of people whose full cure (utility index returns to 1.0) would be equivalent to saving a life (gaining 1.0). This is reported in column 2. Thus, for example, according to the QWB, curing one person from a 'cough' would increase utility by $1 - 0.74 = 0.26$. Four such cures would increase utility by $4 \times 0.26 = 1.04$ and therefore be equivalent to saving a life. The implausibility of this and the other results in Table 1 casts serious doubt upon the existence of this strong interval property for these two instruments.

Some critics of utility measurement have argued that the entire enterprise will fail if the scale includes death or abbreviate life because,

Table 1 Number Cured Equivalent to Saving One Life—Implied by Two MAU Instruments

State	*Published value of state*	*Number cured equivalent to saving a life (approximately)*
McMaster Health Index Questionnaire (HUI Mark 2)[a]		
Some limitations in physical ability to lift, walk, run, jump or bend	0.870	8
Needing a hearing aid	0.870	8
Having pain or discomfort for a few days in a row every month	0.870	8
Needing mechanical aids to get around, but not needing help from others	0.730	4
Quality of Wellbeing Scale (QWB)		
Stuffy, running nose	0.830	6
Pimples	0.800	5
Lisp	0.763	4
Headache	0.756	4
Spells of feeling upset	0.743	4
Trouble with sleeping	0.743	4
Cough	0.743	4

Note: (a) There is now a HUI Mark 3 instrument (see Furlong et al. 1998).

Source: Nord, Richardson and Macarounas-Kirchmann (1993).

as Carr-Hill (1992) argues, there is a 'quite legitimate refusal' of normal people to rate death on the same scale as health states.[5] While intuitively appealing we know of no evidence to support this position and during the construction of utility weights for both the AQoL 1 and AQoL 2 we did not encounter respondents who refused to trade life for quality of life under any circumstances. This does not, of course, indicate that the TTO is the gold standard technique for utility elicitation. This contentious issue is outside the scope of the present article.[6]

In the largest comparative study of MAU instruments to date, Hawthorne et al. (2001) found little in common in the conceptualisation or construction of five instruments and a relatively low correspondence between the utility scores obtained from 976 survey respondents. The correlation coefficients from this study are reported in Table 2 and are low. The data which were correlated were obtained from instrument scoring algorithms which eliminate the 'noise' which exists in individual data: it is 'averaged out'. The resulting scores from each instrument should, therefore, be identical. Despite this, the highest correlation—between AQoL 1 and the 15D instruments (0.821)—implies that only 67 per cent of variation in one instrument is explained by the other instrument. The lowest correlation—between HUI 3 and EQ5D (0.653)—implies that only 43 per cent of variation in one instrument is explained by the other.

The strength of the correlation is a relatively 'soft' test of validity. The null hypothesis that two instruments both give unbiased estimates of true utility would result in a linear relationship between the instruments which passed through the points (0, 0) and (1, 1); that is, an increase in the value of utility measured by one scale would correspond, on average, with an identical increment measured on the second scale. This result was not obtained by Hawthorne et al. (2001). Rather, two groups of instruments were identified. AQoL 1, HUI 3 and EQ5D gave similar utility scores; the 15D and SF36 (Brazier weights) also gave similar scores. However when instruments in the two groups were compared, the slope of the linear relationship differed by up to 100 per cent—differences in the utilities of health states predicted by instruments in the first group would be double the differences found by the two instruments in the second group. This implies that twice the QALY gain would be estimated using instruments from the first group as compared with instruments in the second group.

The results from the five instruments study imply an unsatisfactory 'state of the art' in MAU construction. Results reflect differences in the assumptions and methods at almost every stage of the construction, including the possibility of simple measurement error, a topic receiving relatively little discussion by economists in the utility measurement literature.

1.3 Aims of the AQoL Project

There were two broad objectives of the AQoL project. These were, first, to advance the state of the art of instrument construction and, second, to create an instrument with increased sensitivity, construct and predictive validity. More specifically, the project sought to create utility instruments where the descriptive system was:

Table 2 Correlations between Instruments

	AQoL 1	*HUI 3*	*15D*	*EQ5D*	*SF36*
HUI 3	0.762				
15D	0.821	0.799			
EQ5D	0.751	0.653	0.760		
SF36	0.733	0.664	0.741	0.725	
Mean	0.767	0.715	0.775	0.722	0.716

Note: $N = 968$. The population includes outpatient and ward patients.

Source: Hawthorne et al. (2001).

- derived using correct psychometric procedures for instrument construction (construct validity);
- sensitive to as much of the full universe of health-related QoL as is practical;
- based upon a description of 'handicap'—problems in a social context—as distinct from a 'within the skin' descriptive system;
- based upon structurally independent dimensions of health; and
- hierarchical, so that the descriptive system could allow redundancy—double counting—within dimensions in order to achieve instrument sensitivity, but with structural independence between the dimensions.

The achievement of these properties for the AQoL 1 is described elsewhere (Hawthorne et al. 1997). The AQoL 2 sought to incorporate five additional elements. These were:

(i) an increase in the sensitivity of the descriptive system in the region of full health and a description which permitted the evaluation of health promotional activities as well as illness cure or alleviation;

(ii) the creation of at least two scaling systems based upon the TTO, as with AQoL 1, and the person trade-off (PTO) scaling methodologies (the appropriate choice of scaling instrument has not been determined in the literature);

(iii) a re-estimation of the utility scores employing techniques to eliminate one possible source of bias in previous methodologies, namely a 'focusing effect';

(iv) the testing and use of 'deliberative weights' which permit and encourage the contemplation of the health states for a significant period of time (1 to 2 weeks) before responding to questions; and

(v) the use of a more flexible two-stage modelling methodology to combine disaggregated dimension scores into an overall utility score for a multi-attribute health state.

2. Methodological Issues in the Construction of an MAU Instrument

2.1 Instrument Construction Theory

The correct psychometric procedures for instrument construction are summarised in Figure 1. To our knowledge AQoL 1 is the only instrument which has fully implemented these procedures.

As shown, instrument construction involves theory, data collection and item analysis. Both AQoL 1 and AQoL 2 were based upon the hypothesis that (dis)utility depends primarily upon the extent of a person's handicap; that is, it depends upon the effects of ill health upon a

Figure 1 Steps in Constructing an MAU Instrument (AQOL 2 Specific Information)

1. Theory of health-related QoL
- Concept (handicap)
- Hypothesised dimensions

2. Item bank
- Literature, eclectic sources
- Filter
- Focus groups (GPs, specialists)
- Linguistic analysis

3. Item selection
- Survey 1: 'Construction' (n = 618)
- Statistical analysis (Structural Equation Modelling)

4. Scaling
- Survey 2 (2 × interview + postal, n = 411)

5. Validation
- Survey 3: AQoL (1) n = 972
- Descriptive system validation
  - – Confirmatory factor analysis
  - – Comparison with non-utility scales

person's capacity to function in a social context. In contrast, the HUI instruments and the 15D incorporate descriptive systems based upon disability, that is, upon 'within the skin' descriptions of the impact of ill health upon a person's capacity to carry out certain functions. Next, dimensions of handicap are hypothesised. An item bank is constructed from the large number of items which describe the hypothesised dimensions. AQoL items were obtained from the literature, from other instruments, from focus groups, directly from the clinicians and from the research team itself. The initial items in the item bank are 'filtered' to eliminate items which are poorly expressed, which contain ambiguous or multiple elements ('aspects' or 'concepts') or which are obviously repetitive.

Final item selection is based upon an analysis of a 'construction survey'. This is a stratified and representative group of respondents who complete all of the items. Statistical analyses identify items which cluster together and the correspondence between these clusters and the hypothesised dimension structure. The final choice of items and dimensions is based upon the interplay of empirical results, the theory and the coherence of the overall instrument.

The resulting instrument is scaled (calibrated). The number of health states described by a multi-attribute descriptive system for health status is too large to obtain utility scores separately for each state. For example, the AQoL 1 utility algorithm consists of 12 items each with four response categories.[7] Consequently, there are 4^{12} (16.8 million) combinations of item responses. Scaling therefore requires the use of a model and a combination rule to estimate the utility of each health state from the item responses and the item utilities which have been separately obtained. To date, MAU instruments have employed simple additive models (weights sum to unity), multiplicative models (weights constrain scores between 1.0 and 0), and econometric models (selected multi-attribute states are regressed upon item responses and the coefficients of the best fitting statistical result become item weights).

Finally, instrument construction should be followed by a series of validation studies. Despite the powerful and misleading connotations of the term 'validated', an instrument is never fully validated in the sense that it is shown to be a gold standard. Rather, evidence is obtained which supports the hypothesis that an instrument produces true values for utility *in a particular context*, a principle which was established in the 1950s (Cronbach and Meehl 1955). This process normally involves a series of comparisons with other instruments and with the property in question (for example, different levels of illness) and the evidence supporting the hypothesis of instrument validity is progressively strengthened by the accumulation of confirmatory results. Despite its youth, AQoL 1 has achieved some outstanding results (Hogan et al. 2001; Sturm et al. 2002; Osborne et al. 2003; Hawthorne et al. 2001). As discussed earlier this weak form of 'validation' is necessary but not sufficient for demonstrating that the scores obtained represent a true index of utility.

2.2 Challenges

The process described involves a number of challenges. First, the descriptive system must convey the same information to the survey respondent in the construction survey and to the patient who subsequently uses it to describe their own health state. For example, a 'within the skin' description of hearing loss may elicit a significant disutility when it is initially scaled but a much smaller score from a hearing-impaired respondent if their social environment permits significant adaptation.

Second, an instrument must have an appropriate level of preference independence. Simplifying, the utility score of an item or dimension should not depend upon the health state described by another item or dimension (see Von Winterfeldt and Edwards 1986 for a discussion of preference independence). Without preference independence it would become necessary to model and scale the interactions. Only Feeny et al. (1996) have attempted a partial modelling of such an interaction in the context of the HUI 3. However, their study concluded that a simple multiplicative model without interactions outperformed the partial 'multi-linear' model.

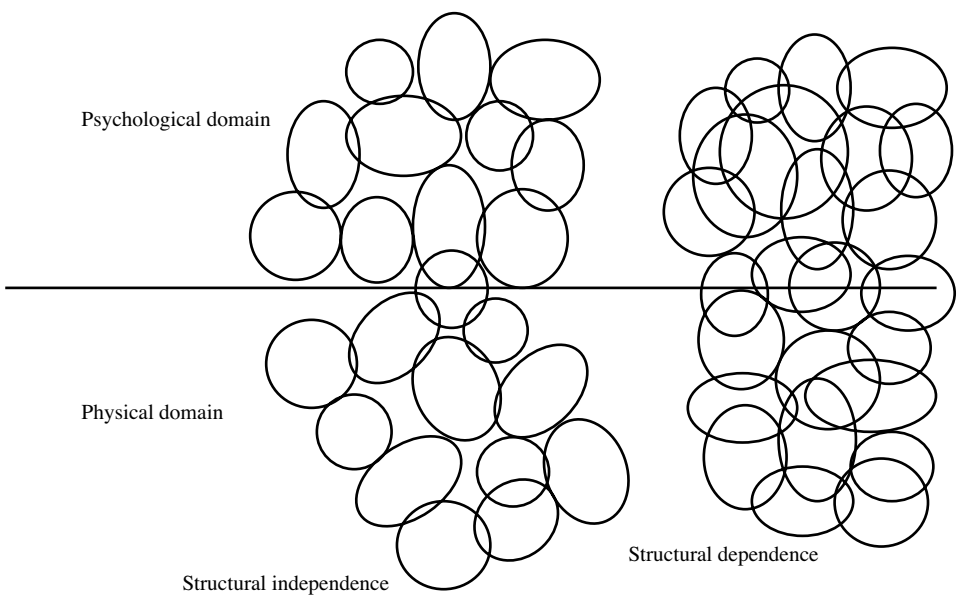
A third requirement is that the items are sensitive to all health states over the health domain which the instrument purports to describe. For example, an instrument which included the disutility from reduced locomotion might accurately detect a reduced capacity to walk and run but fail to detect the reduced capacity to climb stairs. If the former problem did not correlate highly with the latter then the descriptive system will have a degree of insensitivity. More importantly, neither of these problems might have a significant effect upon an elderly person who does not seek to walk significant distances and does not have stairs in their house. The more relevant question might therefore concern the elderly person's ability to carry out the activities of daily living which should, therefore, be included in a sensitive instrument. This example illustrates one of the reasons for basing a descriptive system upon the concept of handicap.

Fourth, and potentially in conflict with the need for instrument sensitivity, the descriptive system should have structural independence—orthogonality—between items or dimensions. In the terminology of decision analysis, there should not be 'redundancy' in the instrument. This will occur if more than one item describes part of an attribute. For example, an instrument separately describing a reduced capacity to walk, to carry out activities of daily living, and to engage in sport and social activities might capture the same problem three different ways. With most forms of scaling this would result in an erroneously low score for individuals with poor mobility.

The trade-off between instrument sensitivity is illustrated in Figure 2, in which the content of an item is represented by an oval. The instrument illustrated on the left-hand side of the figure is close to the ideal structure. Most of the psychological and physical domains are described. Some insensitivity exists where items do not cover parts of the domain. In contrast, the instrument on the right-hand side is sensitive but includes very significant overlap as the majority of points in both of the domains are in more than one oval.

AQoL 1 sought to overcome the latter problems and the trade-off between redundancy and instrument sensitivity by adopting, for the first time, a hierarchical structure. This is shown in Figure 3 in which the manifest items cluster into five latent variables, each representing a dimension of the global latent variable, namely health-related quality of life.

Figure 2 Structural Dependence and Double Counting



Sensitivity within dimensions was sought by employing several items in the knowledge that this resulted in some redundancy within dimensions. Orthogonality was achieved between the five dimensions during the construction stage through the use of factor analysis. The downward bias resulting from double counting was limited by independently assessing the disutility of each dimension 'all worst' health state; that is, it was not possible for the disutility, including redundancy, to be greater than the disutility of the three items evaluated simultaneously.

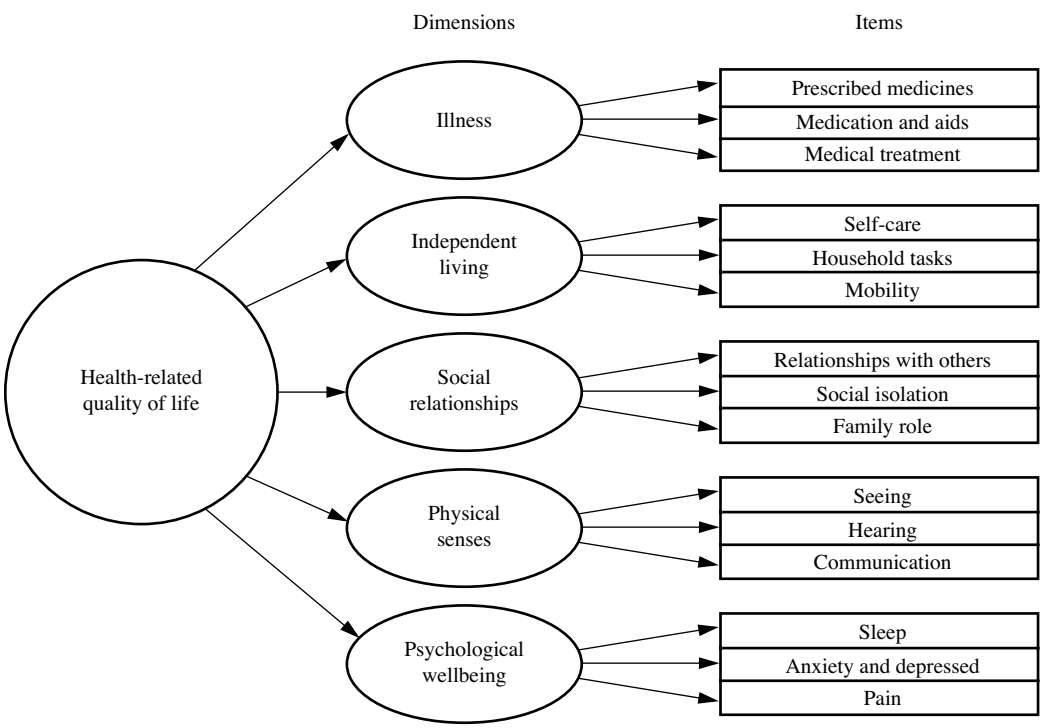
AQoL 2 introduced a somewhat different statistical strategy. Structural Equation Modelling was used to select items and dimensions which maximised the models explanatory power of the variance and covariance between items and dimensions; that is, we selected the model where the latent variable for QoL best explains the co-variance between manifest item responses. This strategy does not, however, ensure orthogonality between dimensions. To offset the effects of redundancy a second stage 'correction' to the magnitude of the predicted utilities will be carried out as discussed in Section 5 below.

2.3 AQoL 2 Innovations

The challenges for AQoL 2 largely arose from the experience with AQoL 1. In the three years following its initial publication, AQoL 1 was requested and sent to 80 research teams and appears to have been used in at least 50 projects. Results from these and from the authors' five-instrument study indicate a number of strengths and weaknesses in AQoL 1 (Hawthorne et al. 2001). Positive features appear to be as follows:

- AQoL 1 has greater sensitivity over certain domains of ill health than other instruments and particularly in the region of good health;
- conceptualising health in terms of handicap has led to a preference for the AQoL instrument in a number of projects where social context has been of importance;

Figure 3 Structure of AQoL 1



- the instrument detects—predicts—greater changes in utility than other instruments included in the comparative studies; and

- the instrument is quickly completed and easy to administer.

Negative features emerging from this experience were as follows:

- Despite the relative sensitivity in the region of good health there is significant room for improvement. Like other instruments, AQoL 1 is primarily concerned with ill health, not vitality and wellness, as needed for the evaluation of health promotional activities.

- Utility scores have been modelled in the AQoL using the most flexible algorithm to date, namely a multi-level multiplicative model. While there are compelling reasons for *preferring* a multiplicative to an additive model there are no reasons for believing that the true structure is a precise, simple, multiplicative relationship between all of the constituent items and dimensions. Further, while items were selected to minimise preference dependency there is no procedure for offsetting bias introduced by this or the other model-induced threats to numerical validity.

- Despite the use of the hierarchical structure to quarantine the effects of structural dependence within dimensions, global AQoL scores are systematically lower than scores on other instruments. This may be, in part, because the AQoL includes in its scoring algorithm provision for the social dimension of people's lives. This is largely excluded by the EQ5D and the HUI 3. However the HUI 3, the other multiplicative model, also has low utility scores which suggests the possibility that the multiplicative model, by permitting lower scores, may tolerate downward errors in a way which cannot occur with additive models.

- As with all other instruments, AQoL 1 employed 'spontaneous utilities'; that is, respondents were presented with TTO questions in an interview context where, despite the exhortation to think about the task, the opportunities for deliberation were small and the opportunities for discussion, consultation and contemplation, non-existent. Because of adaptation, we hypothesised that 'deliberative utilities' would be systematically higher than 'spontaneous utilities'.[8]

- A particular threat to the validity of a decomposed, then reconstructed, instrument score is the so-called 'focusing fallacy' (Ubel et al. 2001). Survey respondents are asked, for example, to consider the disutility of a dimension 'all worst' health state while all other dimensions are at the dimension 'all best'. This 'swing weight' approach to the elicitation of utilities recommended in textbooks on decision analysis facilitates both the process of questioning and the subsequent modelling. In principle, these weights produce an unbiased estimate of the dimension importance, uncontaminated by other aspects of a person's health state. However, the process will yield invalid scores if respondents forget or discount the fact that all other dimensions of health are excellent and consequently they focus exclusively upon the single dimension of poor health and, wrongly, interpret it as indicating overall health, including other dimensions. For example, a respondent who is asked to rate life in a wheelchair may easily forget that with good communication, friendship, no pain and good health in all other respects, and a social environment which allows them to be relatively independent, it is possible to enjoy a relatively good life. If respondents are not reminded of this, the focusing fallacy could create a significant downward bias in estimated utilities.

3. Modelling AQoL 2

As described above, MAU theory requires the initial decomposition of a multi-attribute state into its constituent attributes, their evaluation and subsequent recombination. AQoL 2 has two levels of disaggregation. First, aggregate health states are decomposed into dimensions. Second, dimensions are disaggregated into

items. The recombination at each level requires item and dimension importance weights. Multi-attribute theory suggests that, when the sum of importance weights exceeds unity, a multiplicative model should be used. For independent reasons, this model is also important in the context of health state utilities.[9]

The procedures adopted for the derivation of the AQoL 2 descriptive system followed the psychometric principles outlined above. As noted, AQoL appears to be unique amongst MAU models in this respect. AQoL 2 is similar to AQoL 1 in its conceptualising health primarily in terms of handicap. AQoL 2 was also constructed to achieve a multi-level structure with a number of sub-dimensions, each of which consists of a number of (non-orthogonal) items.

The content of an instrument is determined by the 'universe' of health states defined by the item bank. For AQoL 2, the item bank was expanded to include items of greater relevance in the region of normal to good health. Additionally, response categories for items in AQoL 1 were expanded from four per item in order to increase upper end sensitivity.

The second innovation with respect to the descriptive system was the addition of a 10-point rating scale with endpoints 'greatly improved' and 'totally ruined' (the respondent's life). Respondents were asked to use this scale to indicate how the health state described by their item response affected their QoL. The scale was included for two reasons. First, it permits a consistency check. Discordance between the item response and the rating scale may signal the need to eliminate the response from an analysis. Second, dimension scores may be compared, econometrically, with both item and rating scale responses to determine whether or not the rating scale responses increase the explanatory power of the dimension score. These options are not pursued here. Systematic rules or algorithms for the inclusion of information from the rating scale have yet to be investigated and, at present, rating scale data would need to be used with care.

The protocol for scaling AQoL 2 included three potentially important innovations. The first of these was an attempt to encourage respondent deliberation. The almost universal practice in CUA has been to commence the interview with a brief introduction and 'warm up' exercise and then to present respondents with a vignette or health state and ask for their response (using the TTO or SG). While respondents are encouraged to think before responding, the time constraints upon the interview necessarily result in a 'spontaneous response'. People making real world decisions with respect to these health states would, in contrast, have the opportunity to contemplate the options at length and to discuss the issues with family and friends.

There has been almost no experimentation with the use of 'deliberative responses'. (For exceptions see Murray and Lopez 1996 and Shiell et al. 2000.) Consequently the AQoL 2 protocol employed two separate face-to-face interviews. In the first, the usual protocol was adopted. Interviews were preceded by an introduction and warm up exercise followed by the TTO elicitation. The warm up typically took about 10 minutes but in some cases longer. Respondents were then dichotomised randomly. One-half of respondents were provided with a deliberation kit designed to encourage thought and discussion of selected issues between the interviews (the intervention group). The remaining respondents were simply re-interviewed (the control group). Differences (a low test–retest correlation) between the first and second interview responses in the second group may arise because of the sensitising effect of the first interview (Cook and Campbell 1978) or because of unreliability. Significant differences between the two group's responses to the second interview may be attributed to deliberation. Results presented below employ Stage 2 interview results. Comparison of these with Stage 1 and a comparison of the intervention and control groups is reported in Peacock et al. (2003).

The second and potentially most important innovation in scaling AQoL 2 was a change in the presentation of questions to minimise error arising from the focusing effect. For each of the multi-attribute health states, an overview of the full health state was included which indicated which of the dimensions were at the dimension

all-best, all-worst, or at an intermediate health state. This took the form of a visual aid. When a respondent was asked to focus upon poor health in one dimension only, they were provided this information pictorially in a way which reminded them that other dimensions were good or at their all best.

A final difference with AQoL 1 arises for pragmatic reasons. If a single respondent was asked to provide all of the information required, the interview burden would have been excessive, even allowing for a two-stage interview. Consequently, the two face-to-face interviews were used to collect relatively complex TTO and PTO scores for the major parameters, namely the multi-attribute health states and the dimension all-worst scores. Item responses and item worst scores were collected subsequently from the respondents using a postal survey and a rating scale. Repetition of some rating scale questions during the interview allowed the construction of an econometric 'exchange rate' between rating scale and TTO/PTO scores.

As discussed, to increase the flexibility of the modelling a two-stage procedure was adopted, described below as the 'Stage 1 multiplicative model' and the 'Stage 2 econometric correction'. Stage 1 employed the standard multiplicative model recommended in decision analytic theory (Von Winterfeldt and Edwards 1986). This is similar to equation (1) below:

$$U = U_1 * U_2 * U_3 \ldots * U_n \tag{1}$$

where U is the utility of the combined multi-attribute health state and U_i $(i = 1, \ldots, n)$ are the utility scores for items (in the dimension model) or dimensions (in the AQoL model). The actual model is somewhat more flexible. It is calculated using disutilities rather than utilities and these are adjusted for the relative importance of each of the model's dimensions. This results in equation (2) in which x_{ij} are dimension (or item) scores, w_i are the dimension (or item) importance weights and k is the overall scaling constant. This is obtained by solving equation (3) for k. It is similar to the requirement in an additive model that the dimension weights sum to unity. The relationship between utility and disutility is given in equation (4).

$$DU = \frac{1}{k}\left\{\prod_{i=1}^{n}[1 + kw_i DU_i(x_{ij})] - 1\right\} \tag{2}$$

$$k = \prod_{i=1}^{n}(1 + kw_i) - 1 \tag{3}$$

$$U^* = 1 - DU^* \tag{4}$$

where DU is the disutility score corresponding with utility U.

This multiplicative model was applied at two levels; first, to combine items into dimensions and, second, to combine dimensions into the overall AQoL score.

In order to carry out the 'Stage 2 econometric correction', TTO scores were collected for a selection of MAU health states. These were selected using an experimental design in order to include varying response levels from each of the dimensions and with varying combinations of response levels from the dimensions. These multi-attribute scores have been regressed upon the multiplicative Stage 1 AQoL score and other Stage 1 data using the power function in equation (5).

$$TTO(MA) = AQoL^{a + D} \tag{5}$$

where $AQoL$ = Stage 1 multiplicative AQoL score; a = constant; and D = a set of parameters including dimension scores and slope dummy variables.

This function is constrained to pass through the points (1, 1) and (0, 0); that is, when the multiplicative AQoL score is 0 the predicted score for the power function is 0 and, likewise, an AQoL score of 1.00 must predict a score of 1.00. Between these points the function is flexible and may vary with the model parameters. Results are not included in the present article.

4. Results

4.1 Survey Results

Two postal surveys and two interviews (with the same respondent) were conducted. To achieve a broadly representative sample of the Australian population, names were selected from postcodes within Melbourne according to

Table 3 Data Collection for AQoL 2

Purpose	*Respondents (number)*	*Response rate (per cent)*
Postal Survey 1		
Postal, outpatients, inpatients completion of items in item bank	618[a]	44
Interview 1		
TTO values for dimension worst, multi-attribute health states	411	47
Interview 2		
Multi-attribute health states (continued) PTO, self-TTO	411	47
Postal Survey 2		
Rating scale: Item responses, item worst scores	163	40

Note: (a) General population 316, outpatients 96, inpatients 206.

the socio-economic profile of the postcode, as measured by the SEIFA index. Respondents to the first postal survey—the 'construction survey'—were asked to complete all of the items in the item bank which survived the initial filter. The second postal survey was for the calibration of item responses for items selected for the final instrument. The protocol required the transformation of the rating scale into TTO scores. Details of this procedure and a full description of the surveys and interviews are given in Richardson et al. (2003b). The two interviews were carried out to obtain TTO, PTO and self-TTO data. (The latter two datasets are not discussed here.) A small payment was made to respondents who came to a central location for the two interviews. The number of respondents and response rates are shown in Table 3. Partly because of the persistence with which contacts were pursued and partly because of the financial inducement, response rates for such an onerous interview/survey were acceptable and greater than the 25 to 30 per cent response rate commonly obtained from such surveys.

4.2 AQoL 2 Descriptive System

Structural Equation Modelling was used to determine the dimensions and the combination of items within each dimension which best explained variation in the observed item responses. 'Logical analysis'—correspondence between items, dimensions and theoretical expectations—was also used when the statistical results were ambiguous or perverse.

The result of the analysis is shown in Figure 4 and the AQoL 2 questionnaire is reproduced in Appendix 1. The 20 items selected form six dimensions of health: independent living (four items); social and family (three items); mental health (four items); coping (three items); pain (three items); and sense perceptions (three items). The coefficients reported in Figure 4 indicate an exceptionally good relationship between the postulated model and the pattern of item responses. A confirmatory fit index (CFI) above 0.9 is considered to be acceptable. The CFI of 0.99 for AQoL 2 is very good.[10] Commencing from the left side of Figure 4, the first column of numbers are the gamma coefficients between the dimension and AQoL latent variables. These are equivalent to a standardised correlation coefficient. In contrast with the correlations reported in Table 2 these correlations are based upon individual observations. There is no 'averaging' of the noise and, consequently, such correlation coefficients are generally low. In the present case, however, with the exception of sense perceptions where the gamma coefficient is 0.51, all of the coefficients are 0.73 or greater. Lambda weights between the observed item responses and the dimension latent variables—the middle column of Figure 4—may also be interpreted as equivalent to correlation coefficients. None is below 0.50. Error terms on the individual items in the final, right-hand column are generally low for an analysis of individual-level data.

4.3 Utility Weights

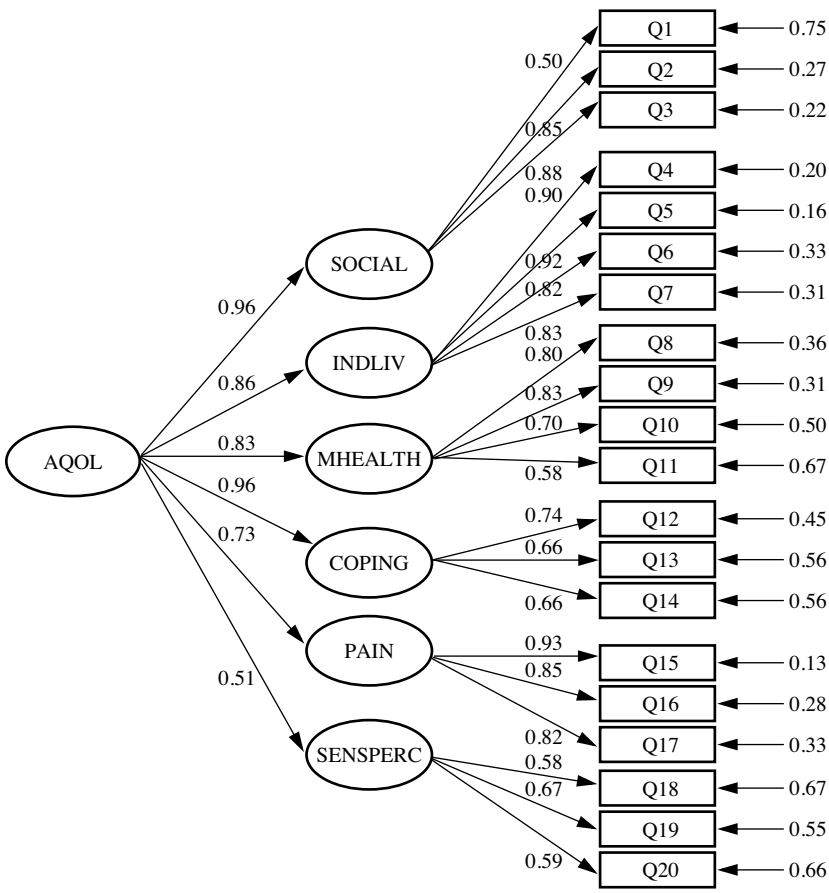
Results from the scaling interviews and postal survey are reported in Tables 4 to 7. Postal survey 2 obtained ratings scale results for item

responses and item worst health states. These were transformed into TTO equivalent scores using a two-part transformation function described in Richardson et al. (2003b). Table 4 reports item response utilities measured on a (1–0) scale where the endpoints are the item best (DU = 0.00) and the item worst (DU = 1.00).

Items were constructed to achieve two objectives. The first was to obtain item responses that are approximately equidistant between the item best and worst health state. Thus, for example, if the disutility scores for an item assumed values of 0.00, 0.02, 0.04, 0.06 and 1.00, then the item would be unable to detect changes in the range 0.06 to 1.00. The second objective was to obtain greater sensitivity near full health than has been achieved in previous instruments. Results in Table 4 indicate that these two objectives have been largely fulfilled. Only four of the 20 items have a space between response values which exceed 0.5 (items 6, 15, 16 and 17). In contrast, the space between the first two response items (which are in the vicinity of good health) is comparatively small. In 12 cases it is less than 0.10 and the maximum gap is 0.20 (items 7 and 16).

Item worst scores, w_i, were also estimated from rating scale results in the second postal survey and are measured on a scale from dimension best (DU = 0.00) to dimension worst (DU = 1.00). They indicate the relative

Figure 4 Structure of AQoL 2[a]



Note: (a) Chi-square = 460.73, df = 164, P-value = 0.00000, RMSEA = 0.054, CFI = 0.99, d1_6×6×2.spl. From the left, the three sets of numbers represent gamma coefficients (between the AQoL and dimension latent variables), Lambda coefficients (between dimension and items) and error terms on each of the items.

importance of the different items. To obtain the final item weights these are multiplied by the dimension scaling factors (k_d) which are derived from the item worst scores and from equation (3) (see Table 5). The final item weight, wt_i, is used to construct the dimension formulae shown later (see Figure 5).

The appropriate criterion for judging the results in Table 5 is that weights should not be too small—indicating an unimportant item—and, ideally, there should be no item in a dimension which dominates other results. From Table 5 these objectives have been achieved. No item has an importance weight of less than 0.38 and 15 of the 20 weights exceed 0.5. No single item dominates the results.

TTO values for the dimension worst and AQoL all-worst health states were assessed on a best health (0.00) – death (1.00) scale. The latter endpoint was used in preference to the AQoL all-worst health state to minimise the cognitive burden upon interviewees. As the all-worst health state may be (and generally was) worse than death for respondents, the TTO protocol permitted this option.[11] The disutilities of these pivotal results were collected in both of the face-to-face interviews which were conducted between two and four weeks apart.

Table 4 Item Disutilities (Mean TTO Scores)

Response level	*Dimension 1*	*Dimension 2*	*Dimension 3*	*Dimension 4*	*Dimension 5*	*Dimension 6*
	Item 1	*Item 5*	*Item 8*	*Item 12*	*Item 15*	*Item 18*
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.07	0.07	0.13	0.06	0.13	0.03
3	0.44	0.46	0.39	0.34	0.64	0.22
4	0.82	0.84	0.84	0.72	1.00	0.62
5	1.00	1.00	1.00	1.00		0.84
6						1.00
	Item 2	*Item 6*	*Item 9*	*Item 13*	*Item 16*	*Item 19*
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.03	0.19	0.14	0.06	0.20	0.02
3	0.24	0.76	0.39	0.38	0.76	0.20
4	0.47	1.00	0.82	0.77	1.00	0.59
5	0.84		1.00	1.00		0.83
6	1.00					1.00
	Item 3	*Item 7*	*Item 10*	*Item 14*	*Item 17*	*Item 20*
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.04	0.20	0.10	0.06	0.07	0.19
3	0.25	0.65	0.33	0.42	0.33	0.70
4	0.57	1.00	0.78	0.83	0.75	1.00
5	0.83		1.00	1.00	1.00	
6	1.00					
	Item 4		*Item 11*			
1	0.00		0.00			
2	0.04		0.06			
3	0.30		0.37			
4	0.80		0.84			
5	1.00		1.00			

Note: Item best and worst disutilities are set equal to 0.00 and 1.00 respectively.

Mean values, reported in Table 6, reveal imperfect test–retest reliability with the second results generally lower than the first, suggesting that, after deliberation, health states appear somewhat less serious than when they are first contemplated. The reported median scores are consistent with mean values and particularly those from the second interview.

The theoretically more plausible 'deliberative' results from the second survey were used in the reported results here in preference to spontaneous weights. Evidence presented by Shiell et al. (2000) suggests that utilities elicited from a second interview are likely to reflect stable future values.

The calculation of dimension weights is shown in Table 7. In this table, w_d are the dimension all-worst TTO scores. The AQoL scaling constant, k, is derived from these six weights and from equation (3). The product of the dimension all-worst scores and the scaling constant give the effective dimension weights wt_d. These are used to derive the overall AQoL formula below.

4.4 *Recalibration to the Best Health – Death Scale*

Insertion of the item and dimension weights into equation (2) produces disutility scores

Table 5 Item Weights for Use in Dimension Models

Dimension / *Item*	$(-)\ k_d * w_i = wt_i$					*Dimension* / *Item*	$(-)\ k_d * w_i = wt_i$				
Independent living						Coping					
1	(0.978)	*	(0.39)	=	0.38	12	(0.930)	*	(0.42)	=	0.39
2	(0.978)	*	(0.59)	=	0.58	13	(0.930)	*	(0.64)	=	0.60
3	(0.978)	*	(0.63)	=	0.62	14	(0.930)	*	(0.77)	=	0.72
4	(0.978)	*	(0.80)	=	0.78						
Social and family						Pain					
5	(0.923)	*	(0.64)	=	0.59	15	(0.962)	*	(0.63)	=	0.61
6	(0.923)	*	(0.70)	=	0.65	16	(0.962)	*	(0.77)	=	0.74
7	(0.923)	*	(0.51)	=	0.47	17	(0.962)	*	(0.65)	=	0.57
Mental health						Sensory					
8	(0.983)	*	(0.64)	=	0.63	18	(0.851)	*	(0.58)	=	0.49
9	(0.983)	*	(0.59)	=	0.58	19	(0.851)	*	(0.46)	=	0.39
10	(0.983)	*	(0.65)	=	0.64	20	(0.851)	*	(0.61)	=	0.52
11	(0.983)	*	(0.71)	=	0.70						

Note: k_d = dimension scaling constants; w_i = item worst *DU*; and wt_i = item weight.

Table 6 Dimension Worst Disutility Scores

	Interview 1		*Interview 2*				
Dimension	*Mean*	*Standard error*	*Mean*	*Standard error*	*Median*	*Per cent negative*	*Number*
1. Independent living	0.54	(0.02)	0.47	(0.02)	0.40	8.5	367
2. Social and family	0.50	(0.02)	0.45	(0.02)	0.40	6.3	367
3. Mental health	0.51	(0.02)	0.48	(0.02)	0.45	7.4	367
4. Coping	0.35	(0.02)	0.35	(0.02)	0.30	1.1	367
5. Pain	0.54	(0.02)	0.59	(0.02)	0.50	16.1	367
6. Sensory perception	0.68	(0.02)	0.64	(0.02)	0.60	19.1	367

Table 7 Dimension Weights for Use in AQoL Model

Dimension	k	*	w_d	=	wt_d
1. Independent living	0.965	*	(0.47)	=	0.454
2. Social	0.965	*	(0.45)	=	0.434
3. Mental health	0.965	*	(0.48)	=	0.463
4. Coping	0.965	*	(0.35)	=	0.338
5. Pain	0.965	*	(0.59)	=	0.570
6. Senses	0.965	*	(0.64)	=	0.618
AQoL	W/k	=	1.132/ 0.965	=	1.17

Note: k = AQoL scaling constant; w_d = dimension all-worst; wt_d = dimension weight; and W = AQoL all-worst (full health – death scale).

constrained to the range 0.00 to 1.00. For the overall AQoL model these endpoints correspond with the AQoL best and worst health state respectively. To recalibrate to an AQoL best health – death scale, where death equals 1.00, requires the multiplication of the (0–1) model scores by W, the disutility of the AQoL all-worst health state measured on a life–death (0–1) scale. This latter value was 1.102.

The final dimension and overall AQoL formulae are obtained by inserting the item and dimension weights from Tables 5 and 7 into equation (1), rescaling to the full health – death scale as described above and converting disutility into utility using equation (4). Results are presented in Figure 5.

An example of the use of these formulae to obtain a health state utility score is given in Appendix 2.

5. Discussion and Future Work

There are numerous unresolved issues associated with the construction of MAU instruments and, more fundamentally, with the measurement and the valuation of the outcomes from health-related interventions. As decisions concerning patient treatment and the net benefit of different services for a health scheme are being made daily, it is important that these are based upon current best practice. This rationale extends to the construction of instruments for measuring the quality of life.

Figure 5 Multiplicative Utility Formulae

General formula for utility model[a]

$$U_d = \frac{1}{k_d} \cdot \prod_{i=1}^{n} [1 - k_d w_i (1 - u_i)] - \left(\frac{1}{k_d} - 1\right)$$

Independent living $U_1 = 1.02_i[(0.62 + 0.38u_1)(0.42 + 0.58u_2)(0.38 + 0.62du)(0.22 + 0.78u_4)] - 0.02$

Social and family $U_2 = 1.08_i[(0.41 + 0.59u_5)(0.36 + 0.64u_6)(0.53 + 0.47u_7)] - 0.08$

Mental health $U_3 = 1.02_i[(0.37 + 0.63u_8)(0.42 + 0.58u_9)(0.36 + 0.64u_{10})(0.30 + 0.70u_{11})] - 0.02$

Coping $U_4 = 1.08_i[(0.61 + 0.39u_{12})(0.41 + 0.59u_{13})(0.28 + 0.72u_{14})] - 0.08$

Pain $U_5 = 1.04_i[(0.39 + 0.61u_{15})(0.26 + 0.74u_{16})(0.37 + 0.63u_{17})] - 0.04$

Senses $U_6 = 1.18_i[(0.51 + 0.49u_{18})(0.61 + 0.39u_{19})(0.49 + 0.51u_{20})] - 0.18$

AQoL general formula

$$U_{AQoL} = \frac{W}{k} \cdot \prod_{d=1}^{n} [1 - kw_d (1 - U_d)] - W\left(\frac{1}{k} - 1\right)$$

$$U_{AQoL} = 1.17[(0.546 + 0.454u_1)(0.566 + 0.434u_2)(0.537 + 0.463u_3)(0.662 + 0.338u_4)(0.430 + 0.570u_5)(0.382 + 0.618u_6)] - 0.17$$

Note: (a) The utility formula is derived from equations (2) and (4) earlier. U_d = utility score, dimension d; u_i = utility score, item i; k_d = scaling constant, dimension d; k = scaling constant, AQoL; w_i = item worst, item i; and w_d = dimension worst, dimension d.

Despite this, it is important to recognise the limitations of current state-of-the-art instrument construction and to progressively improve measurement. Some of these limitations motivated the present study and have been addressed in the present article. These include the need for deliberation before values are elicited from respondents and an interview protocol which explicitly overcomes the focusing effect. Technical 'validity' has been increased by using standard psychometric procedures for constructing a model in the form currently used in the MAU literature and by the introduction of multi-level modelling as a superior methodology for explaining variance and covariance between the manifest item responses.

During the two-stage interviews, data were collected which allow two other methodological developments. While most utility measurements now employ the TTO or SG (which produce very similar results) the PTO has also been advocated. In particular, the PTO is probably the preferred scaling instrument if a 'community' perspective is desired—the PTO asks respondents for a judgement concerning the allocation of health between others, and in a way that leaves the respondent personally unaffected. The technique is important as it was the procedure used in the calculation of disability-adjusted life years (DALYs) in the World Health Organization (WHO) Global Burden of Disease Study which estimates the DALY burden of every disease in every country (Murray and Lopez 1996). The procedure has been promoted vigorously through WHO workshops and it has been widely adopted in economic evaluation studies and, in particular, in Australia. The choice between the personal perspective of the TTO and the social perspective of the PTO is an ethical, and not a technical, issue. Consequently, PTO scores have been collected and a PTO version of the AQoL algorithm will be forthcoming.

Modelling in the MAU literature has been relatively unsophisticated. The multiplicative model represents a significant improvement upon the additive model. However, it is still a relatively simple combination rule which imposes a questionable degree of uniformity in the relationships between items and dimensions: there is no difference in this relationship in the vicinity of full health and in the vicinity of death. The same multiplicative relationship is assumed to exist between all items and dimensions. As there is no theoretical rationale for this uniformity a second stage adjustment is being developed to introduce flexibility into both of these relationships. This involves the econometric 'explanation' of TTO scores for a representative sample of multi-attribute health states, using survey data and, in principle, any other relevant patient characteristics as the independent variables. The econometric relationship which best explains the multi-attribute TTO scores will be the corrected generic AQoL instrument.

It was noted earlier that the multiplicative models in the literature—the HUI 1–3 and AQoL 1—produce lower utility scores than the other models. An additional reason for the second stage was to correct any such downward bias arising from the multiplicative model.

To date, the more than 1000 TTO scores obtained for selected multi-attribute health states have been used as dependent variables in an econometric analysis which has employed a modified power function which constrains the function to pass through the pivotal points (0, 0) and (1, 1) (see Richardson et al. 2003b). If the Stage 1 multiplicative model explained all of the systematic variation in the utility of health states then the resulting 'power function' would have an exponent of 1.00; that is, the function would be the linear relationship $U = U_{AQoL}$. If the multiplicative model does not explain all of the systematic variation then the function will be more complex. In principle, any variable which improves the functional relationship might be included in the formula. The interval property of the resulting utility scores would be promoted, not confounded, by the transformation. The key assumption here—and in the relevant literature—is that this property applies to the utility of the multi-attribute health states (the left-hand side of the equation). The multiplicative approximation (right-hand side) will not have this property unless it is identical to the MAU scores. The

independent variables used to date in this analysis have been the predicted AQoL score from the multiplicative model, dummy variables for the quartile of the (0–1) range of this score and six variables repeating the dimension scores used in the Stage 1 model. Results from the analysis have been encouraging with R^2 coefficients between 0.67 and 0.76 (in equations where the constant term has been suppressed). Importantly, this 'second stage correction' may be used to obtain low-cost adaptations of the AQoL for atypical diseases or population groups with an atypical preference structure.

6. Conclusions

While we have attempted to improve upon the current methodologies, the work reported in this article does not, of course, guarantee validity and reliability. This question remains problematical as no one has devised a test for gold standard validation. This would require, inter alia, demonstration that utility weights accurately represented the desired trade-off between the quality and quantity of life. There has been widespread acceptance in the MAU literature that the gold standard should employ community-stated preferences. However, this view is contestable. Elsewhere in economics, consumer (revealed) preferences are usually sovereign. Current practice might be rationalised by arguing that community preferences represent patient preferences. This is not universally true as patients with long-term health problems undergo significant adaptation to their health state. The extent of the divergence between adapted and non-adapted preferences and its relevance have not been properly documented or discussed (Menzel et al. 2002).

Even with the more tractable concept of community preferences, there are unresolved theoretical and measurement issues. With either concept, benefits are conceptualised as an increased quality or quantity of life. Neither concept addresses the numerous other procedural and distributive elements of health-related social wellbeing which authors have posited in the recent literature. Nord (1999) and Nord et al. (1999) have suggested that the term 'Cost Value Analysis' be used to indicate a shift from individual utility as the relevant metric in economic evaluation studies to 'social value' as measured by a QALY adjusted to include presently neglected elements of social wellbeing.

The chief justification for the present generation of measurement methods is that they are an improvement upon the methods of the past and that the systematic inclusion of an increasing number of socially desirable attributes and the use of increasingly sophisticated methodologies will result in better decision making. It is almost self-evident, however, that these methodologies are incomplete and should be viewed as an aid to decision making and not as a definitive algorithm for social choice.

First version received July 2003;
final version accepted December 2003 (Eds).

Appendix 1: Assessment of Quality of Life (AQoL) Mark 2

How to answer

Please read the Explanatory Statement and sign a consent form before you begin.

Each question has two parts. You answer the first part by ticking the box next to the response that best fits your situation. The second part of each question is a horizontal scale. You mark a cross somewhere along the scale to show how your quality of life is affected by the situation you describe in your answer to the first part of the question. Look at the example answer for more information.

Example answer

Mr Smith's relationships with his family make him generally happy, so he marks the second box from the top to show his answer:

i) My relationships with my **family** make me:
- ❑ very happy
- ☑ generally happy
- ❑ neither happy nor unhappy
- ❑ generally unhappy
- ❑ very unhappy
- ❑ this question is not relevant to me.

Mr Smith feels his quality-of-life is greatly improved by the fact that his relationships with his family make him 'generally unhappy', so he marks a cross on the left hand end of the scale.

How does this affect my quality of life?

		×								

greatly improved — no effect either way — totally ruined

When you finish answering all the questions, please hand the questionnaire back.

Many thanks!

Assessment of Quality of Life (AQoL) Mark 2

Q1 How much help do I need with household tasks (e.g. preparing food, cleaning the house or gardening)?
- ❑ I can do all these tasks very quickly and efficiently without any help
- ❑ I can do these tasks relatively easily without help
- ❑ I can do these tasks only very slowly without help
- ❑ I cannot do most of these tasks unless I have help
- ❑ I can do none of these tasks by myself.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q2 Thinking about how easy or difficult it is for me to get around by myself outside my house (e.g. shopping, visiting):
- ❑ getting around is enjoyable and easy
- ❑ I have no difficulty getting around outside my house
- ❑ a little difficulty
- ❑ moderate difficulty
- ❑ a lot of difficulty
- ❑ I cannot get around unless somebody is there to help me.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q3 Thinking about how well I can walk:
- ❑ I find walking or running very easy
- ❑ I have no real difficulty with walking or running
- ❑ I find walking or running slightly difficult. I cannot run to catch a tram or train, I find walking uphill difficult.
- ❑ walking is difficult for me. I walk short distances only, I have difficulty walking up stairs.
- ❑ I have great difficulty walking. I cannot walk without a walking stick or frame, or someone to help me.
- ❑ I am bedridden.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q4 Thinking about washing myself, toileting, dressing, eating or looking after my appearance:

- ❑ these tasks are very easy for me
- ❑ I have no real difficulty in carrying out these tasks
- ❑ I find some of these tasks difficult, but I manage to do them on my own
- ❑ many of these tasks are difficult, and I need help to do them
- ❑ I cannot do these tasks by myself at all.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q5 My close and intimate relationships (including any sexual relationships) make me:

- ❑ very happy
- ❑ generally happy
- ❑ neither happy nor unhappy
- ❑ generally unhappy
- ❑ very unhappy.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q6 Thinking about my health and my relationship with my family:

- ❑ my role in the family is unaffected by my health
- ❑ there are some parts of my family role I cannot carry out
- ❑ there are many parts of my family role I cannot carry out
- ❑ I cannot carry out any part of my family role.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q7 Thinking about my health and my role in my community (that is to say neighbourhood, sporting, work, church or cultural groups):

- ❑ my role in the community is unaffected by my health
- ❑ there are some parts of my community role I cannot carry out
- ❑ there are many parts of my community role I cannot carry out
- ❑ I cannot carry out any part of my community role.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q8 How often did I feel in despair over the last seven days?

- ❑ never
- ❑ occasionally
- ❑ sometimes
- ❑ often
- ❑ all the time.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q9 And still thinking about the last seven days: how often did I feel worried?

- ❑ never
- ❑ occasionally
- ❑ sometimes
- ❑ often
- ❑ all the time.

How does this affect my quality of life?

greatly improved — no effect either way — totally ruined

Q10 How often do I feel sad?

- ❑ never
- ❑ rarely
- ❑ some of the time
- ❑ usually
- ❑ nearly all the time.

How does this affect my quality of life?

greatly improved	no effect either way	totally ruined

Q11 When I think about whether I am calm and tranquil or agitated:
- ❑ always calm and tranquil
- ❑ usually calm and tranquil
- ❑ sometimes calm and tranquil, sometimes agitated
- ❑ usually agitated
- ❑ always agitated.

How does this affect my quality of life?

greatly improved	no effect either way	totally ruined

Q12 Thinking about how much energy I have to do the things I want to do, I am:
- ❑ always full of energy
- ❑ usually full of energy
- ❑ occasionally energetic
- ❑ usually tired and lacking energy
- ❑ always tired and lacking energy.

How does this affect my quality of life?

greatly improved	no effect either way	totally ruined

Q13 How often do I feel in control of my life?
- ❑ always
- ❑ mostly
- ❑ sometimes
- ❑ only occasionally
- ❑ never.

How does this affect my quality of life?

greatly improved	no effect either way	totally ruined

Q14 How much do I feel I can cope with life's problems?
- ❑ completely
- ❑ mostly
- ❑ partly
- ❑ very little
- ❑ not at all.

How does this affect my quality of life?

greatly improved	no effect either way	totally ruined

Q15 Thinking about how often I experience **serious pain**. I experience it:
- ❑ very rarely
- ❑ less than once a week
- ❑ three to four times a week
- ❑ most of the time.

How does this affect my quality of life?

greatly improved	no effect either way	totally ruined

Q16 How much pain or discomfort do I experience?
- ❑ none at all
- ❑ I have moderate pain
- ❑ I suffer from severe pain
- ❑ I suffer unbearable pain.

How does this affect my quality of life?

greatly improved	no effect either way	totally ruined

Q17 How often does pain interfere with my usual activities?
- ❑ never
- ❑ rarely
- ❑ sometimes
- ❑ often
- ❑ always.

How does this affect my quality of life?

greatly improved					no effect either way					totally ruined

Q18 Thinking about my vision (using my glasses or contact lenses if needed):

- ❑ I have excellent sight
- ❑ I see normally
- ❑ I have some difficulty focusing on things, or I do not see them sharply. E.g. small print, a newspaper or seeing objects in the distance.
- ❑ I have a lot of difficulty seeing things. My vision is blurred. I can see just enough to get by with.
- ❑ I only see general shapes. I need a guide to move around.
- ❑ I am completely blind.

How does this affect my quality of life?

greatly improved					no effect either way					totally ruined

Q19 Thinking about my hearing (using my hearing aid if needed):

- ❑ I have excellent hearing
- ❑ I hear normally
- ❑ I have some difficulty hearing or I do not hear clearly. I have trouble hearing softly-spoken people or when there is background noise.
- ❑ I have difficulty hearing things clearly. Often I do not understand what is said. I usually do not take part in conversations because I cannot hear what is said.
- ❑ I hear very little indeed. I cannot fully understand loud voices speaking directly to me.
- ❑ I am completely deaf.

How does this affect my quality of life?

greatly improved					no effect either way					totally ruined

Q20 When I communicate with others, e.g. by talking, listening, writing or signing:

- ❑ I have no trouble speaking to them or understanding what they are saying
- ❑ I have some difficulty being understood by people who do not know me. I have no trouble understanding what others are saying to me.
- ❑ I am understood only by people who know me well. I have great trouble understanding what others are saying to me.
- ❑ I cannot adequately communicate with others.

How does this affect my quality of life?

greatly improved					no effect either way					totally ruined

AQoL Study Background Questions

Please tick ✓ one box per question.

21 Are you:
❑ male ❑ female

22 In what year were you born? 19____

23 Where were you born?
❑ Australia ❑ Other country
☞ Which one? _____

24 Is English your first language?
❑ yes ❑ no ☞ Specify: _____

25 What is your highest level of education?
❑ primary schooling only
❑ secondary schooling completed
❑ secondary schooling not completed.
☞ How many years completed? ____
❑ trade qualification or TAFE:
☞ Specify course: _____
❑ University or other tertiary study
❑ Other or not applicable: please describe:
_____

26 Which best describes your work situation? (Tick as many boxes as apply)
❑ full-time: self-employed or employee

❑ part-time: self-employed or employee
❑ unemployed, seeking work
❑ working in the home / home duties
❑ retired
❑ student
❑ other: please describe:

_____

If You Are Employed Or Self-Employed Or Seeking Work:

27 What is your occupation?

_____

28 What do you do in your job?

_____

29 Do you receive any Government pension or benefit?
❑ no
❑ yes ☞ Which pension(s) or benefit(s)?

_____

30 Are you:
❑ married or living with a partner
❑ single: never married
❑ single: widowed
❑ single: divorced or separated

31 How would you rate your current level of health, for someone of your age?
❑ excellent
❑ very good
❑ good
❑ fair
❑ poor
❑ very poor
❑ extremely poor

32 Mark one box on the scale to show how important or unimportant religion or spirituality is in your life

❑ ❑ ❑ ❑ ❑ ❑ ❑
⇓ ⇓ ⇓ ⇓
very important important unimportant very unimportant

33 Please mark one box to show your HOUSEHOLD income, either annually, monthly or weekly. Include income that comes to the household from all sources. You may estimate either before or after tax.

❑	yearly	under $20,000
	monthly	under $1,665
	fortnightly	under $800
	weekly	under $385
❑	yearly	$20,001–$30,000
	monthly	$1,665–$2,500
	fortnightly	$800–$1,155
	weekly	$385–$575
❑	yearly	$30,001–$40,000
	monthly	$2,501–$3,330
	fortnightly	$1,156–$1,535
	weekly	$576–$770
❑	yearly	$40,001–$50,000
	monthly	$3,331–$4,165
	fortnightly	$1,536–$1,925
	weekly	$771–$960
❑	yearly	$50,001–$60,000
	monthly	$4,166–$5,000
	fortnightly	$1,926–$2,305
	weekly	$961–$1,155
❑	yearly	$60,001–$80,000
	monthly	$5,001–$6665
	fortnightly	$2,306–$3,075
	weekly	$1,156–$1,540
❑	yearly	more than $80,000
	monthly	more than $6665
	fortnightly	more than $3,075
	weekly	more than $1,540

34 Please mark a box to show whether your answer is before or after tax.
❑ before tax
❑ after tax

Thank you! Please bring this questionnaire with you when you attend the group session/interview.

Appendix 2: Obtaining a Health State Utility from the AQoL Algorithm: A Worked Example

Obtaining a utility score for a health state involves the following steps.

(i) Complete the AQoL questionnaire and determine the 20 response levels which define the health state.

(ii) Read the 20 item disutility scores, du_i, which correspond with the response levels from Table 4. These 'disutilities' are measured on a (1–0) scale with the item best and worst defining the endpoints.

(iii) Enter the item disutility scores, du_i, into the corresponding equation in Figure 5. Calculate the six dimension disutility scores DU_d. These disutilities are measured on a (0–1) scale where the endpoints are the dimension best and dimension 'all worst' (all items at their worst level).

(iv) Enter the six dimension DU_d scores into the final AQoL equation in Figure 5. The score obtained is the predicted disutility for the health state.

(v) Convert disutilities into utilities using the equation $U = 1 - DU$.

These steps are illustrated for a randomly chosen health state in Figure A1.

Endnotes

1. TTO and SG techniques are described in Torrance (1986) and Richardson (1991). The TTO quantifies preferences by asking the number of years (of a given maximum) which would be sacrificed by an individual to be in best health rather than in the health state being measured. The SG is similar but uses the probability of death rather than the number of years sacrificed as the device for eliciting preferences.

2. In principle every health state may be individually measured. In practice, the number of health states in the descriptive system is normally so large that this is infeasible. The only example of this approach is the original Rosser

Figure A1 Calculating a Utility Score: A Numerical Example

1. Complete the AQoL questionnaire to obtain 20 response levels; 1 per item

   Example: Response levels are:

   $D1(1, 1, 2, 1)$; $D2(2, 2, 1)$; $D3(3, 2, 3, 1)$; $D4(1, 1, 1)$; $D5(2, 1, 1)$; $D6(2, 1, 2)$

2. Read the 20 *disutility* scores DU from Table 4

   $D1(0, 0, 0.04, 0)$; $D2(0.07, 0.19, 0)$; $D3(0.39, 0.14, 0.33, 0)$; $D4(0, 0, 0)$; $D5(0.13, 0, 0)$; $D6(0.03, 0, 0.19)$

3. Convert to utilities $U = 1 - DU$

   $D1(1.0, 1.0, 0.96, 1.0)$; $D2(0.93, 0.81, 1.0)$; $D3(0.61, 0.86, 0.67, 1.0)$
   $D4(1.0, 1.0, 1.0)$; $D5(0.87, 1.0, 1.0)$; $D6(0.97, 1.0, 0.81)$

4. Enter the 20 utility scores into the relevant equation in Figure 5 and calculate dimension scores

   Dimension utilities

   $U_1 = 1.02_i\{[0.62 + 0.38(1.0)][0.42 + 0.58(1.0)][0.38 + 0.62(0.96)][0.22 + 0.78(1.0)]\} - 0.02 = 0.975$

   $U_2 = 1.08_i\{[0.41 + 0.59(0.93)][0.36 + 0.64(0.81)][0.53 + 0.47(1.0)]\} - 0.08 = 0.829$

   $U_3 = 1.02_i\{[0.37 + 0.63(0.61)][0.42 + 0.58(0.86)][0.36 + 0.64(0.67)][0.30 + 0.70(1.0)]\} - 0.02 = 0.54$

   $U_4 = 1.08_i\{[0.61 + 0.39(1.0)][0.41 + 0.59(1.0)][0.28 + 0.72(1.0)]\} - 0.08 = 1.00$

   $U_5 = 1.04_i\{[0.39 + 0.61(0.87)][0.26 + 0.74(1.0)][0.37 + 0.63(1.0)]\} - 0.04 = 0.918$

   $U_6 = 1.18_i\{[0.51 + 0.49(0.97)][0.61 + 0.39(1.0)][0.49 + 0.51(0.8)]\} - 0.18 = 0.864$

5. Enter the six dimension scores into the AQoL formula in Figure 5 and calculate the health state utility

   Health state utility

   $U_{AQoL} = 1.17\{[0.546 + 0.454(0.975)][0.566 + 0.434(0.829)][0.537 + 0.463(0.54)]$
   $[0.662 + 0.338(1.0)][0.430 + 0.570(0.918)][0.382 + 0.618(0.864)]\} - 0.17 = 0.569$

Kind Index which is now seldom used because of its limited sensitivity.

3. These three approaches have been used respectively to construct the EQ5D (EuroQoL) (Williams 1995), the Quality of Wellbeing (QWB) (Kaplan et al. 1996), the 15D (Sintonen and Pekurinen 1993), the Health Utilities Index (HUI) 1–3 (Feeny, Torrance and Furlong 1996), and AQoL 1 (Hawthorne et al. 1997).

4. See Richardson (1994) for a critique of the rating scale as a method for eliciting utility values. In recent literature the rating scale has been largely abandoned in this context.

5. We are grateful to an anonymous reviewer for drawing our attention to this reference.

6. There is a long list of such contentious issues. For example, there is a clear preference in the literature for eliciting utilities from the general population. It is, however, possible to argue that it is the preferences of patients or potential patients which should be elicited. This *ethical* question and numerous others are reviewed in Brazier et al. (1999).

7. The initial instrument consisted of 15 such items but the three items describing 'illness' were removed as a result of the validation study.

8. This problem is distinct from the issue of adaptation which occurs when people alter their lives, their expectations or their values in a way which accommodates their poor health states. This issue is highly contentious and includes, inter alia, the question of whose preferences should be elicited, the general public, unadapted patients or patients following adaptation. As with all MAU instruments, AQoL 1 and AQoL 2 employed the preferences of the general public. It is likely that different results would be obtained for all instruments (and other utility studies) if post-adaptation preferences of patients were adopted. This issue is outside the scope of the present article but is discussed more fully in Menzel et al. (2002).

9. It is possible for a number of health states to independently impact catastrophically upon the QoL. For example, intense pain and intense depression may both reduce the QoL to zero. This cannot be described in a simple additive model where importance weights must sum to unity and, consequently, the importance weights on depression and upon pain must be numerically small. In contrast, the multiplicative model permits any dimension to reduce QoL to a level equivalent to death.

10. The Chi-squared value for the model is 460.73 with 164 degrees of freedom. The associated probability is close to zero, indicating that, according to the Chi-squared test, our model is very unlikely to represent the *exact* value of the data. However, the Chi-squared is recognised as a misleading statistic in this context as the hypothesis of interest is that the model is a good (if not perfect) fit. Thus, as the sample size increases, the probability that the model is a perfect fit decreases for any value of the Chi-squared. An analogy is to consider the relationship $Y = 0.99X$ which models the true relationship $Y = X$. With a sufficient sample size we could reject with any level of confidence the hypothesis that $Y = X$ but the model, nevertheless, provides a very good approximation. This limitation on the Chi-squared measure is documented in the literature and is noted in the LISREL program documentation where it is stated that 'since Chi-squared N – 1 times the minimum value of the fit function, Chi-squared tends to be large in large samples if the model does not hold. A number of goodness of fit measures have been proposed to eliminate or reduce its dependence on sample size' (LISREL Help File, Goodness of Fit Indices).

11. Negative TTO scores are obtained by offering the option of immediate death or the health state of interest for n years followed by full health for the remainder of the person's life. If death is preferred the implied value of the health state is negative. Calculating disutilities from these results is problematical as there is no lower limit to the implied negative 'utility'. The problem is discussed in Richardson and Hawthorne (2001).

References

Brazier, J., Deverill, M., Green, C., Harper, R. and Booth, A. 1999, 'A review of the use of health status measures in economic evaluation', *Health Technology Assessment*, vol. 3, no. 9, pp. 1–154.

Carr-Hill, R. A. 1992, 'Health related quality of life measurement—Euro style', *Health Policy*, vol. 20, pp. 321–8.

Cook, T. D. and Campbell, D. T. 1978, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Miffin, Boston.

Cronbach, J. and Meehl, P. 1955, 'Construct validity in psychological tests', *Psychological Bulletin*, vol. 52, pp. 281–302.

Feeny, D., Torrance, G. and Furlong, W. 1996, 'Health Utilities Index', in *Quality of Life and Pharmacoeconomics in Clinical Trials*, ed. B. Spilker, Lippincott Raven Publishers, Philadelphia.

Furlong, W., Feeny, D., Torrance, G. et al. 1998, 'Multiplicative multi-attribute utility function for the Health Utilities Index Mark 3 (HUI3) system: A technical report', Working Paper 98-11, Centre for Health Economics and Policy Analysis, Hamilton McMaster University.

Hawthorne, G., Richardson, J. and Day, N. A. 2001, 'A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments', *Annals of Medicine*, vol. 33, pp. 358–70.

Hawthorne, G., Richardson, J., Osborne, R. and McNeil, H. 1997, 'The Assessment of Quality of Life (AQoL) instrument: Construction, initial validation and utility scaling', Working Paper 76, Centre for Health Program Evaluation, Monash University.

Hogan, A., Hawthorne, G., Kethel, L. et al. 2001, 'Health-related quality-of-life outcomes from adult cochlear implantation: A cross-sectional survey', *Cochlear Implants International*, vol. 2, pp. 115–28.

Kaplan, R., Ganiats, T., Sieber, W. and Anderson, J. P. 1996, 'The quality of wellbeing scale', *Medical Outcomes Trust Bulletin*, vol. 4, pp. 2–3.

Menzel, P., Dolan, P., Richardson, J. and Olsen, J. A. 2002, 'The role of adaptation to disability and disease in health state valuation: A preliminary normative analysis', *Social Science and Medicine*, vol. 55, pp. 2149–58.

Murray, C. and Lopez, A. 1996, *The Global Burden of Disease*, Harvard School of Public Health on behalf of World Health Organization and World Bank, Cambridge, Massachusetts.

Nord, E. 1999, *Cost Value Analysis*, Cambridge University Press, Cambridge.

Nord, E., Pinto, J.-L., Richardson, J., Menzel, P. and Ubel, P. 1999, 'Incorporating societal concerns for fairness in numerical valuations of health programmes', *Health Economics*, vol. 8, pp. 25–39.

Nord, E., Richardson, J. and Macarounas-Kirchmann, K. 1993, 'Social evaluation of health care versus personal evaluation of health states: Evidence on the validity of four health-state scaling instruments using Norwegian and Australian survey data', *International Journal of Technology Assessment in Health Care*, vol. 9, pp. 463–78.

Osborne, R. H., Hawthorne, G., Lew, E. A. and Gray, L. C. 2003, 'Quality of life assessment in the community-dwelling elderly: Validation of the Assessment of Quality of Life (AQoL) instrument and comparison with the SF-36', *Journal of Clinical Epidemiology*, vol. 56, pp. 138–47.

Peacock, S., Richardson, J., Hawthorne, G., Day, N. A. and Iezzi, A. 2003, 'The Assessment of Quality of Life (AQoL) 2 instrument: The effect of deliberation and alternative utility weights in a multi attribute utility instrument', Working Paper 143, Health Economics Unit, Monash University, <http://heu.buseco.monash.edu.au>.

Richardson, J. 1991, 'Economic assessment of health care: Theory in practice', *Australian Economic Review*, 1st quarter, pp. 4–21.

Richardson, J. 1994, 'Cost utility analysis: What should be measured?', *Social Science and Medicine*, vol. 39, pp. 7–21.

Richardson, J. 2002, 'The conceptual basis for summary measures of population health: Qualifying DALYs; dallying with QALYs', in *Summary Measures of Population Health: Concepts, Ethics, Measurement and*

Applications, eds C. J. L. Murray, J. A. Salomon, C. D. Mathers and A. D. Lopez, World Health Organization, Geneva.

Richardson, J. and Hawthorne, G. 2001, 'Negative utility scores and evaluating the AQoL all worst health state', Working Paper 113, Centre for Health Program Evaluation, Monash University.

Richardson, J., Hawthorne, G., Day, N. A., Peacock, S. and Iezzi, A. 2003a, 'The Assessment of Quality of Life (AQoL) II instrument: Conceptualising the Assessment of Quality of Life instrument Mark 2 (AQoL 2) methodological innovations and the development of the AQoL descriptive system', Working Paper 141, Health Economics Unit, Monash University, <http://heu.buseco.monash.edu.au>.

Richardson, J., Peacock, S., Day, N. A., Hawthorne, G. and Iezzi, A. 2003b, 'The Assessment of Quality of Life (AQoL) 2 instrument: Derivation of the scaling weights using a multiplicative model and econometric second stage correction', Working Paper 142, Health Economics Unit, Monash University, <http://heu.buseco.monash.edu.au>.

Richardson, J., Peacock, S., Hawthorne, G., Day, N. A. and Iezzi, A. 2003c, 'The Assessment of Quality of Life (AQoL) 2 instrument: Overview of the Assessment of Quality of Life Mark 2 project', Working Paper 144, Health Economics Unit, Monash University, <http://heu.buseco.monash.edu.au>.

Shiell, A., Seymour, J., Hawe, P. and Cameron, S. 2000. 'Are preferences over health states complete?', *Health Economics*, vol. 9, pp. 47–55.

Sintonen, H. and Pekurinen, M. 1993, 'A fifteen dimensional measure of health related quality of life (15D) and its applications', in *Quality of Life Assessment*, eds S. Walker and R. Rosser, Kluwer Academic Publishers, Dordrecht.

Sturm, J. W., Osborne, R. H., Dewey, H. M., Donnan, G. A., Macdonell, R. A. L. and Thrift, A. G. 2002, 'Brief comprehensive quality of life assessment after stroke: The Assessment of Quality of Life instrument in the North East Melbourne Stroke Incidence Study (NEMESIS)', *Stroke*, vol. 33, pp. 2888–94.

Torrance, G. 1986, 'Measurement of health state utilities for economic appraisal: A review', *Journal of Health Economics*, vol. 5, pp. 1–30.

Ubel, P. A., Loewenstein, G., Hershey, J., Baron, J., Mohr, T., Asch, D. A. and Jepson, C. 2001, 'Do nonpatients underestimate the quality of life associated with chronic health conditions because of a focusing illusion?', *Medical Decision Making*, vol. 21, pp. 190–9.

Von Winterfeldt, D. V. and Edwards, W. 1986, *Decision Analysis and Behavioural Research*, Cambridge University Press, Cambridge.

Williams, A. 1995, 'The measurement and validation of health: A chronicle', Discussion Paper 136, Centre for Health Economics, University of York.