MONASH University
Business and Economics

Centre for Health Economics

Research Paper 2007 (24)

Construction and Validation of the Assessment of Quality of Life (AQoL) Mark II Instrument

Professor Jeff Richardson
Foundation Director, Centre for Health Economics,
Monash University

Dr Stuart Peacock
Director, Centre for Health Economics in Cancer,
British Columbia Cancer Agency

Mr Angelo Iezzi
Research Fellow, Centre for Health Economics,
Monash University

Mr Neil Atherton Day
Principal Research Fellow, Centre for Program Evaluation,
The University of Melbourne

Associate Professor Graeme Hawthorne
Principal Research Fellow, Department of Psychiatry,
The University of Melbourne

November, 2007

ISSN 1833 – 1173
ISBN 1 921187 23 9

Correspondence:

Jeff Richardson
Centre for Health Economics
Faculty of Business and Economics
Building 75
Monash University Victoria 3800
Australia

Phone: +61 (0)3 9905 0754, Fax: +61 (0)3 9905 8344
jeff.richardson@buseco.monash.edu.au

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF BOXES

ABSTRACT

The AQoL program explored methodological innovations in MAU instrument construction using psychometric methods to develop a descriptive system based upon handicap and the use of a multi level descriptive system in which latent dimension variables are constructed from individual items and which, in turn, create the latent AQoL variable. The paper reports chief results and discusses problems associated with negative utilities, spontaneous versus deliberative weights, framing effect bias. New validation procedures are used to test for the existence of a 'strong interval property', the defining characteristic of the QALY. One external validation test of the strong interval property is reported.

Construction and Validation of the Assessment of Quality of Life (AQoL) Mark II Instrument:

1. Background

Multi-attribute utility (MAU) instruments seek to measure the utility of health states in a way which is suitable for use in economic evaluation studies and, particularly, cost utility analysis (CUA). In principle, they have an advantage over holistic health state valuation – a one-off evaluation of an entire health state vignette – as they describe a large number of health states and may be used with a relatively low cost to a research budget. In practice, because of the cost and complexity of constructing MAU instruments, only a limited number have been created to date and recently the literature has been dominated by the use of three of these, namely the EQ5D, SF6D and the HUI III. Two of the earlier instruments, the 15D and QWB are also in use[1].

The present article outlines the construction of another instrument, the Assessment of Quality of Life (AQoL) Mark 2, the objectives it sought to achieve and the extent to which it did this. The article outlines the construction of the descriptive instrument, its 3 stage scaling (calibration) and its validation. Each stage includes a number of new procedures.

The construction of a MAU instrument involves four distinct steps. First, a generic 'descriptive system' must be created. This involves the decomposition of the concept of a health state into multiple dimensions – attributes – of health, each of which may be described by one or more items – questions with multiple responses. Secondly, these items are 'scaled'- scored – in such a way that a single item score may be obtained. Third, the item scores must be combined using some form of modelling to achieve the single score. Fourth, some form of 'validation' must be undertaken: determining whether the instrument's predicted values pass some tests which suggest that they have the properties of utility scores.

A final instrument may be assessed according to several criteria. First, is it valid and reliable; do the final scores represent utilities and are the same scores obtained when the instrument is used with the same patients more than once? Secondly, is the instrument sensitive: do the changes in health states result in changes in the utilities predicted by the instrument? Third, is the instrument easy to use; can it be administered quickly and with little difficulty?

Each stage of the construction has alternative methodologies and has involved varying levels of controversy. Health states may be described a number of ways with a major difference between 'within the skin' conceptualisation (impairment, disability of body functions) and 'handicap' (inability to function in a social context). Descriptive systems may be constructed ad hoc or following psychometric principles of instrument construction. Various scaling methods may be used; time trade-off (TTO), standard gamble (SG), rating scales (RS), person trade-off (PTO) and each of these have been used in MAU instruments. Items may be combined using additive, multiplicative or econometric models. Finally, instrument 'validation' commonly involves tests which justify less confidence in the instrument than is implied by the compelling term 'validated' which connotes the universality of test results which is generally not warranted. A typical 'validation' test consists of the demonstration of a correlation between instrument scores and

[1] Instruments are described in the following references. HUI 3: Feeny et al (1996), Furlong et al (2001), Horseman et al (2003). EQ5D: Kind (1996), Dolan et al (1995), EuroQoL Group (2001). QWB: Kaplan et al (1996a; 1988). 15D: Sintonen and Pekurinen (1993), Sintonen (2001). SF6D Brazier et al (1998), Brazier et al (1999), Brazier et al (2002).

those of another 'validated' instrument which is commonly disease specific and does not purport to measure utility. Not only are such tests context specific but they provide no evidence of the 'strong interval' property which is the defining characteristics of the concept of utility used in the Quality Adjusted Life Year (QALY).[2]

The dominating feature of comparisons reported in Tables 1 and 2 is the lack of similarity between instruments. In principle, this does not imply the invalidity of comparisons based upon different instruments. Just as weight may be measured using spring, balance or electronic scales, different models might result in similar results. Nevertheless, the extent of the differences suggest that different instruments might have different comparative advantage, with the measurement of different health states and that different outcomes from different instruments would be unsurprising.

Table 1. Properties of the major utility instruments

Scale	Coverage* (a)	Type of description[1] (b)	No of dimensions	Valuing method[2]	Psychometric properties		Combination model	Instrument boundaries[4]
					Construct[3]	Validation		
Rosser-Kind	XX	Impairment	2	ME	No	No	None	–1.49 — 1.00
QWB	X	Impairment/ disability	4	VAS	No	Yes	Additive	0.00 — 1.00
15D	✓✓	Impairment	15	VAS	No	Yes	Additive	+0.11 — 1.00
HUI I	X	Impairment	4	TTO	No	No	Multiplicative	–0.21 — 1.00
HUI II	✓	Impairment/ disability	7	VAS/SG	No	Yes	Multiplicative	–0.03 — 1.00
HUI III	✓✓	Impairment/ disability	8	VAS/SG	No	Yes	Multiplicative	–0.36 — 1.00
EQ-5D	X	Handicap	5	TTO	No	No	Regression/ Additive	–0.59 — 1.00
DALY	XX	Disease	N/A	PTO	No	No	RS/PTO[5]	N/A
WHOQoL-Bref	✓✓	Handicap	4	N/A	Yes	Yes	Additive	N/A
SF6D	✓✓	Handicap	6	SG	Yes	No	Additive	+0.46 — 1.00
AQoL	✓✓	Handicap	4	TTO	Yes	Yes	Multiplicative	–0.04 — 1.00

Notes:

* = Coverage of the HRQoL universe, as defined by a review of 14 HRQoL instruments, 1971–1993 (Hawthorne and Richardson 1995). Coding scheme: XX = very poor, X = poor, ✓ = good, ✓✓ = very good.

1 = Based on WHO classification of diseases and impairments (WHO 1980).

2 = ME: Magnitude estimation; VAS: Rating Scale; TTO: Time Trade-off; SG: Standard Gamble; PTO: Person Trade-off

3 = Descriptive system constructed following standard psychometric rules for instrument construction (Ware et al. 1993; WHOQoL Group 1998).

4 = Lower and upper boundaries shown where 0.00 = death and 1.00 = full health. Negative values indicate health states worse than death. Lower boundaries determined by the instrument's 'all worst health state'; upper boundaries determined by the 'all best health state'.

5 = Rating scale validated using the PTO

Source: Hawthorne, Richardson and Day (2001) p 359

[2] Richardson (1994) refers to the commonly used concept as a 'weak interval property' and defines the 'strong interval property' as occurring when a given percentage increase in measured quality of life has the same value, in some meaningful sense, as the same percentage increase in the length of life.

Table 2. Dimensions included in five utility instruments*

Dimensions	AQoL	EQ5D	15D	SF6D	HUI III
'Within the Skin'					
(1) Mobility	*	*	*	**	*
(2) Physical ability/vitality	*		**	***	*
(3) Pain	*	*	*	*	*
(4) Senses	**		*****		****
(5) Bodily care	*	*		*	
(6) General health					
(7) Depression/Anxiety	*	*	**	**	
(8) Cognitive ability			*		*
(9) Memory					*
Handicap					
(1) Social function	*			*	
(2) Family role	*				
(3) Work function				*	
(4) Activities of daily living	*	*	*	*	
(5) Communication	*		*		**
(6) Emotion					*
(7) Intimacy/Isolation	*				
(8) Sexual relationships			*		
(9) Medical aids use					
(10) Medical treatment					

*Asterisks indicate the number of items used in calculation of utility scores.
Source: Hawthorne, Richardson and Day (2001) p 363

This concern was realised in the survey of 878 individuals (142 inpatients, 333 hospital outpatients and 403 members of the general public). The overall correlation between five instruments varied from 0.65 (HUI III, EQ5D) to 0.82 (15D, AQoL). Excluding the SF6D which was subsequently rescaled, utility scores for those aged 16-35 in the community varied from 0.84 to 0.92 (AQoL, 15D) and for inpatients over 66 from 0.43 (AQoL) to 0.55 (HUI III) to 0.74 (15D). At the individual level scores of 1.00 on the EQ5D corresponded with scores as low as 0.3 on the HUI III and conversely, scores of 1.00 on the HUI III obtained negative EQ5D scores for some individuals. Of greatest concern was the relationship between incremental changes measured on different instruments as these determined the incremental QALY values after an intervention. Table 3 reports the coefficient 'b' found in the linear relationship $Y = a + bX$ between instrument scores X and Y, calculated from the procedures with which the relationship is unaffected by the choice of dependent and independent variables. Choice of instrument may more than double estimated QALY values.

Table 3. Linear relationship between instruments

	$X_1 = a_1 + a_2 . X_2$ or $X_2 = a_2 + (1/b) X_1$			
	X_2			
X_1	**AQoL**	**HUI**	**EuroQoL**	**15D**
AQoL		1.05	1.13	2.1
HUI	0.95		0.99	2.0
EuroQoL	0.88	1.01		1.9
15D	0.49	0.5	0.5	

Source: Barnett, VD 1969, 'Simultaneous pairwise linear structural relationships', *Biometrics*, 15, 129-142.

The objective of the initial Assessment of Quality of Life (AQoL 1) was the creation of an instrument using the psychometric principles for instrument construction and an instrument with three additional properties, namely increased sensitivity, a descriptive system based upon handicap and structural independence between dimensions. These properties were sought to overcome perceived limitations in existing instruments. EQ5D has a simple and limited descriptive system with significant 'gaps' between item responses and from the omission of sensory perceptions. The more detailed Health Utilities Index (HUI III) has a descriptive system based upon a 'within the skin' classification of health attributes. This is suitable for some health state dimensions but not those where the most important consequences depend upon social context and the person's ability to perform normally or well in this context. While having elements of handicap, the SF6D items are also limited.

The third challenge – structural independence – arises because of the increased likelihood of a non-orthogonality between items and dimensions as the number of items rises, and the increased likelihood of 'double counting' of disutility when elements of health states are imbedded in more than one item. For example, a person experiencing significant pain is also likely to experience a reduction in physical activity and vitality.

To simultaneously achieve these properties, the first AQoL instrument adopted an hierarchical descriptive system in which non-orthogonality was permitted within dimensions when this was necessary to increase sensitivity but which achieved orthogonality between dimensions of its descriptive system through the use of exploratory factor analysis, with varimax rotation (Hawthorne et al. 1999). In principle, the effects of non-orthogonality (redundancy or double counting) would be therefore contained within dimensions and minimised. The instrument was scaled using TTO methods and a two part model in which multiplicative modelling was used firstly to construct five dimension values and secondly, to combine these dimensions to obtain a single AQoL utility score (Hawthorne and Richardson 1997). The large five instrument study was undertaken to compare AQoL 1 with other instruments. As noted a striking feature of the literature reporting the development of MAUs has been the emphasis on ill health states: no MAU instrument to date has been constructed from the perspective of public health.

The AQoL 2 project was undertaken partly in response to unresolved challenges and partly to achieve some additional objectives. These were (i) to increase the sensitivity of AQoL 1 in the range of normal-good health (a domain neglected in most multi-attribute instruments) and to make the instrument additionally suitable for the evaluation of health promotional (lifestyle) interventions; (ii) to increase the sensitivity of AQoL 1 in other domains and to revise its structure with the use of structural equation modelling; (iii) to revise the TTO methodology to reduce the likelihood that a 'focusing effect' might introduce bias, as suggested by Ubel and Lowenstein (2001); (iv) to introduce third stage modelling to 'correct' for the residual effects of redundancy

(structural dependence or double counting); (v) to ensure that incremental changes predicted from the MAU algorithm corresponded with incremental differences in utilities elicited from specific holistic health states; and (vi) to conduct validation tests to ensure that this last property had been achieved.

Construction methods for the AQoL 2 are outlined in Section 2 emphasising elements which are innovative. Section 3 presents results for the descriptive system, scaling survey and the stage 3 econometric adjustment. Section 4 outlines the validation analysis and presents results from it. Use of the instrument is discussed in the concluding Section 5.

2. Methods

The stages of construction of the AQoL 2 are shown in Box 1. Five distinct steps were involved. First, an overall concept of health and its dimensions was selected using the criterion that it should be the concept most closely related to individual utility. Secondly, the concept was operationalised by describing/defining it with a very large number of items. Thirdly, after initial item reduction the final selection of items was made largely (but not entirely) by quantitative analysis of patient/public responses to the item. Fourthly, selected items and dimensions were scaled/calibrated using a time trade-off (TTO) instrument and dimensions/overall AQoL scores were modelled. Finally, initial validation studies were carried out to test the hypothesis that the final model algorithm predicted holistic TTO scores.

Box 1. Steps in constructing a descriptive system

1. **Theory or HRQoL**
   - Concept (of health)
   - Hypothesised dimensions
2. **Item Bank**
   - Literature, eclectic sources
   - Focus groups (doctors, public)
   - Triage
   - Linguistic analysis
3. **Descriptive Instrument (Survey 1)**
   - Item, dimension selection
   - Confirmatory factor(SEM) analysis
4. **Scaling (Survey 2, 3, 4)**
   - Model selection
   - Stage 1: Dimensions
   - Stage 2: AQoL
   - Stage 3: Adjustment
5. **Validation**
   - 'Internal' data analysis descriptive system
   - External data analysis (Survey 5)

The overarching theory behind the AQoL 2 was that utility is primarily determined by a person's capacity to operate satisfactorily in a social context. The theory is derived from the WHO's concept of social handicap as 'a disadvantage for a given individual, resulting from an impairment or disability... (which) limits or prevents the fulfilment of a role that is normal... for that individual' (WHO, 1980 p 29). Thus, for example, the loss of an eye (impairment) may result in disability (eg

an inability to drive) which may result in social handicap (the person may become isolated in their community). In this example social isolation is the primary determinant of the loss of utility: a blind person who has adapted to their circumstances and who used a taxi or other transportation for communication, may suffer little loss of utility. As no classification of health elements is complete, non-handicap elements were permitted in the item bank and instrument.
The item bank was constructed with items adapted from AQoL 1, the literature and eclectic sources. These included multiple items expressing the same simple homogeneous concept (element). The item bank was tested and supplemented through review using focus groups of doctors and the public. Items were subject to linguistic analysis to achieve a uniform style and triaged to select a 'short list' using the triage criteria of clarity, simplicity and the need for several items per element for the statistical analysis.

Following psychometric theory, the final selection of items and dimensions drew upon the results of a 'construction survey' of individuals likely to encounter the health states which the AQoL 2 sought to describe (as distinct from a representative population survey). Items have been commonly selected in instrument construction on the basis of logical and content analysis alone (Bowling 2001, 2005). However responses often reflect idiosyncrasies of language and secondary connotations of words which are not easily or accurately accounted for in logical analysis and where the implication of content overlap with other items or combinations of items may be opaque. Techniques used in psychometric theory are specifically designed to take account of these factors.

AQoL 1 relied primarily upon principal components and exploratory factor analyses to determine firstly, that the AQoL 1 corresponded with a coherent latent variable and, secondly, that the constituent dimensions were orthogonal in the psychometric sense; that is, that all of the common variance between the dimensions was explained by the latent variable and the error variances were uncorrelated. AQoL 2 relied primarily upon confirmatory factor analysis, a sub-set of SEM, to select the best fitting items for the structure initially determined by AQoL 1. Nevertheless dimension structures were tested and varied. As with AQoL 1 the model finally selected imposed (psychometric) orthogonality upon dimensions.

Structural Dependence

The different approaches in AQoL 1 and AQoL 2 were a consequence of the experience with AQoL 1. In the five instrument comparative study reported above it was found that AQoL 1 and the Canadian Health Utilities Index (HUI III) systematically produced lower scores than in other instruments. These were also the two instruments which had used a multiplicative model for combining the disaggregated scores. This approach to modelling has the great advantage of being able to combine a very large number of items and item responses. This contrasts with an econometric model (EQ5D) where the necessary sample size of respondents needed to produce valid TTO scores sharply rises with the number of possible health states. However the results above suggest that AQoL 1 may not have successfully eliminated redundancy, and that the simple multiplicative model may produce higher disutility scores when more opportunities arise for detecting disutility in a health state, ie when additional items are included in the instrument in order to increase sensitivity.

AQoL 1 sought to overcome this dilemma by seeking sensitivity within dimensions, by relaxing the requirement of within dimension orthogonality, but minimising the effect of this by first imposing orthogonality between dimensions and, secondly, by independently measuring the maximum disutility of each dimension and constraining dimension scores to be less than or equal

to this dimension disutility. Thus, redundancy within AQoL 1 dimensions could not result in a disutility greater than the independently measured dimension disutility.

With hindsight, this strategy may have been compromised for two reasons. First, while maximum dimension disutility was constrained, intra-dimensional scores were not, and the cumulative effect of this might have depressed scores. Secondly, dimensions were only orthogonal in the psychometric sense; ie that all of the non random variance was explained by variation in the overall AQoL 1 latent variable. They were, nevertheless, correlated in the usual statistical sense and the theoretical relationship between this type of non-orthogonality and redundancy is unclear, leaving the possibility of overall double counting of disutilities.

A third factor associated with the presentation of the TTO during the scaling of AQoL 1 was the possibility of a 'focusing effect' identified by Ubel and Lowenstein (2001). This is a problem which may arise if the negative, but not the positive, elements of a health state are presented and respondents only focus upon the negative elements. The approach to this problem is described in Appendix 2.

To offset any or all of the potential problems, AQoL 2 adopted a three stage modelling strategy. In the first two parts, the multiplicative modelling of dimension and overall AQoL scores, the requirement of item orthogonality was initially dropped and the criterion for instrument selection became an item's psychometric properties and particularly the extent to which it contributed to the dimension description. As noted, dimensions were subsequently constrained to be psychometrically orthogonal but this was not relied upon to eliminate redundancy. This was achieved in the third stage of the modelling when scores from the multiplicative modelling, were used to econometrically explain TTO values obtained from independently collected multi-attribute health states constructed from the AQoL 2 descriptive system. The final algorithm adjusted the AQoL 2 multiplicative score using results from the best fitting econometric relationship.

Preference Dependence

The third stage econometric procedure described above also mitigated a further potential problem associated with preference dependence (Feeny et al. 1996; von Winterfeldt and Edwards 1986). Stage 1 and 2 modelling employed multiplicative modelling which allowed greater flexibility than simple additive modelling. However models may also be ''multi linear'. The utility value of an item response may depend upon *the level* of a person on another item or dimension[3]. Because of the difficulty of modelling multi linear relationships where every utility score may, potentially, differ with every health state in the descriptive system first order (preference) independence is usually assumed[4]. However the econometric specification of the stage 3 model discussed below permits adjustment if this problem is found to be consistently significant between two dimensions.

[3] For example, a person who suffers mild pain may experience a greater loss of utility from the pain if they are also depressed than if they are exuberant.

[4] The Health Utilities Index 3 instrument tested for a (partial) multi linear relationship but found a multiplicative model to be superior (Feeny et al 2002). While preliminary results indicated quantitatively important interactions amongst attributes the multiplicative model out-performed the part multi linear function.

Scaling AQoL

Four problems were encountered with the collection of TTO weights for the scaling of AQoL. These are discussed in Appendices 2-5. First, as noted above, MAU modelling is potentially subject to a framing effect. Visual props designed to overcome this are illustrated in Appendix 2. Secondly, TTO scores as normally measured may fall to minus infinity and rapidly asymptote to meaninglessness in the negative range. The approach to this is discussed in Appendix 3. Thirdly, because of the limited research budget item response scores were obtained in a postal rating scale survey and transformed into TTO values using a function described in Appendix 4. Finally, and a general problem with utility elicitation techniques, (more or less) 'spontaneous' responses to answers (after warm up examples) may result in greater disutilities than post deliberative results because of the partial 'shock horror' of contemplating an extended period of ill health. A small RCT to test this hypothesis is summarised in Appendix 5.

Modelling

The first two stages of the recombination of items into utility scores employed a multiplicative model. This is similar to equation 1 below.

$$U = \prod_{i=1}^{n} U_i \qquad \ldots (1)$$

where U is the combined multiplicative score and U_i are the item or dimension scores.
The actual model is somewhat more flexible. It was calculated using disutilities rather than utilities and these are adjusted for the relative importance of each of the dimensions or items. This resulted in equation 2 in which w_i are the dimension (or item) weights and k is the overall scaling constant. This was obtained by solving equation 3. It is similar to the requirement in an additive model that the dimension weights sum to unity.

$$DU = \frac{1}{k}\left[\prod_{i=1}^{n}\left[1 + kw_i DU\left(x_i\right)\right] - 1\right] \qquad \ldots (2)$$

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

The relationship between utility and disutility is given in equation 4.

$$U^* = 1 - DU^* \qquad \ldots (4)$$

The model was applied at two levels to combine items into dimensions and, secondly, to combine dimensions into the overall AQoL score.

The multiplicative formula constrains DU scores to a (0.00-1.00) best-worst scale. Consequently to map values upon a (0.00-1.00) life-death utility scale, values must be adjusted according to equation 5 where DU is the final utility score and W is the ratio of (any) health state measured on a life-death (0, 1) scale to the score measured on a best-worst model (0,1) scale. For this transformation the AQoL 2 all-worst health state was used.

$$DU = W . DU^* \qquad \ldots (5)$$

The third stage modelling was based upon the assumption that the TTO score for a holistic (multi-attribute) health state (MA) TTO represents the 'gold standard', ie that the procedure for combining attributes (items and dimensions) should therefore result in the same estimate of utility as obtained by the evaluation of the multi-attribute health state defined by the same attributes. If it is no longer assumed that the multiplicative model is without bias then there is no clear guideline concerning the relationship between the stage 2 AQoL and the gold standard. It cannot be assumed, for example, that the same functional relationship between (MA) TTO and AQoL scores would exist at high levels and at low levels of disutility. The functional relationship could adjust the stage 2 score at least one of four forms: (i) uniform adjustment of the overall score in all dimensions; (ii) adjustment of a single dimension; (iii) correction for redundancy and preference dependence between a combination of two or more dimensions; and (iv) structural change in the relationship at different levels of disutility. This implies the need for what might be described as 'loose cannon' modelling – the empirical exploration of alternative relationships. Only two general constraints exist. First, the functional relationship must produce utility (TTO) scores between 1.00 and 0.00 when AQoL scores have the value 1.00 and 0.00 respectively. These scores correspond with best health and death respectively. Secondly, a positive AQoL increment should result in a positive increase in utility.

To achieve this degree of flexibility the relationship in equation 6 was employed.

$$TTO = AQoL^{x}$$

$$x = \alpha_0 + \sum_j \alpha_i D_i + \sum_i \sum_j B_{ij} D_i D_j + \sum_{i=1}^{4} shift_i \qquad \text{... (6)}$$

Where
- α_0 = constant
- D_i = dimension score for dimension i
- D_iD_j = dimension D_i times Dimension D_j
- $shift_i$ = dummy variables indicating that the TTO has a disutility score in excess of 0.2; 0.4; 0.6; 0.8.

If the multiplicative model has no bias then the coefficient α_0 would equal unity and other coefficients would be insignificant. If α is not equal to 1.00, a uniform exponential adjustment is made to all DU values. D_i coefficient adjusts for any net bias in individual dimensions; D_iD_j coefficients adjust for interaction between dimensions and the shift coefficients included eliminate any residual over or under estimation of the true TTO values at particular levels of utility. The power function relationship ensures that the predicted score passes through the points (0.00, 0.00) and (1.00, 1.00). As the modelling is initially conducted in disutility space these points represent best health and death respectively in the results below. As the exponent may be negative it is possible for a disutility score to be in excess of 1.00, indicating a health state worse than death.

3. Results

Two postal surveys and two sets of interviews were conducted with each respondent. Results are summarised in Table 4. The first postal survey was for the completion of items in the item bank. The two sets of interviews were carried out to obtain TTO data for dimensions and AQoL 2 all worst scores and for a selection of scores for holistic AQoL states for use in the third stage econometric analysis. The second postal survey was to obtain rating scale assessments of item response levels.

The number of respondents and response rates are shown in Table 1.

Table 4. Data collection for AQoL 2

Purpose (Postal) construction sample	**respondents** **n**	**response rate** **%**
Survey 1: (Postal) completion of items in item bank	618	45
Survey 2: (Interview 1):= dimension worsts, multi-attribute health states (TTO scores)	411	47
Survey 3 (Interview 2) dimension worst (repeat) Multi-attribute health states (cont) PTO, self TTO	411	47
Survey 4 Postal Survey 2: Item responses, item worst scores (Rating Scale)	163	40

AQoL 2 Descriptive System

Construction of the descriptive system is described in Richardson et al. (2004a). Econometric modelling is described in Richardson et al. (2004b).

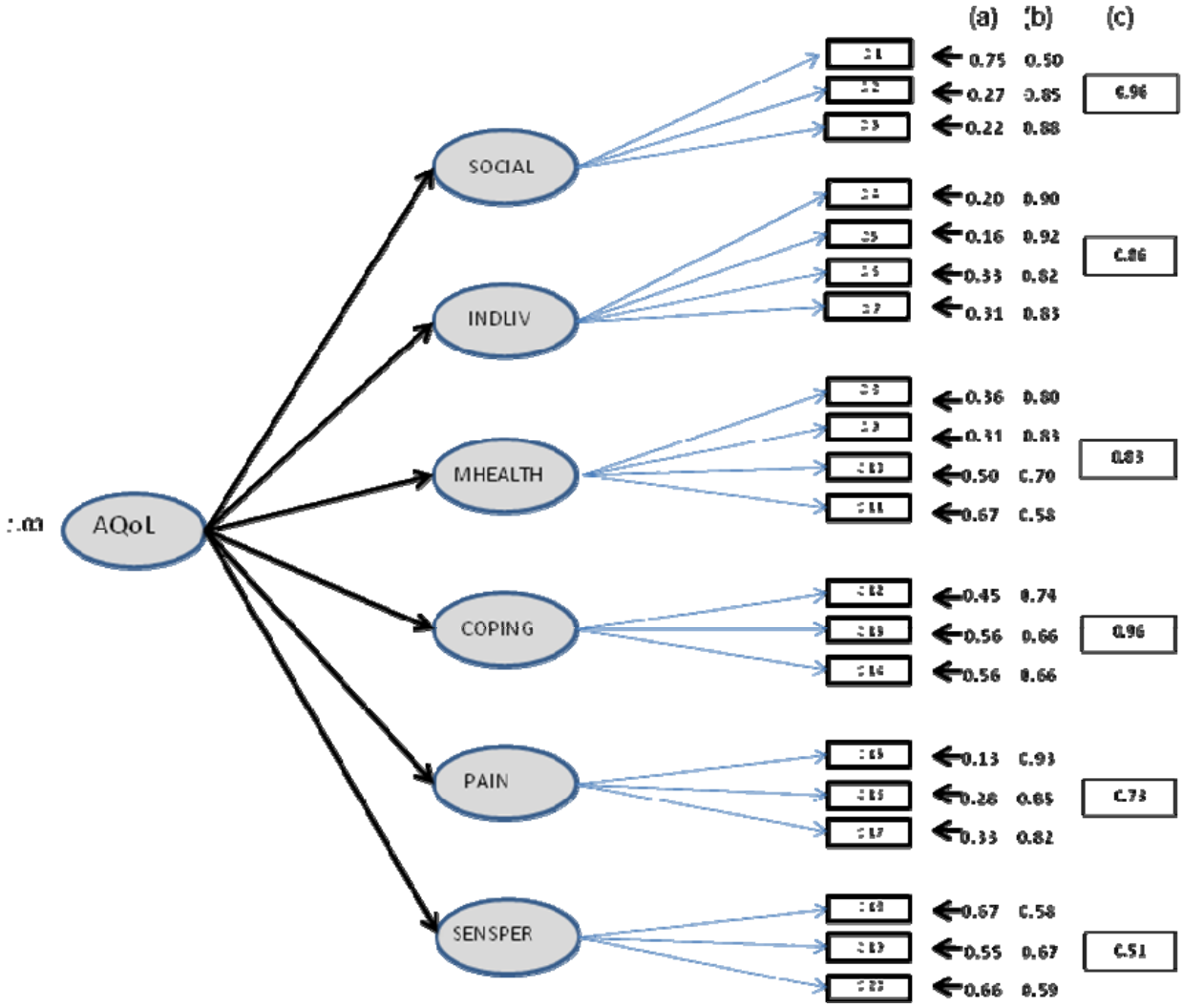
The final construction questionnaire included 112 items and covered 6 domains: (i) Social (including work, family and intimate relationships); (ii) Independent Living; (iv) Mental Health; (iv) Illness (including pain); (v) Values and Beliefs; and (vi) Sensory Items. The survey was administered to 316 randomly selected members of the adult population aged 18 years or above, 96 inpatients and 206 outpatients from Melbourne hospitals. This provided a total of 618 useable responses. The response rate from inpatients was effectively 100 percent; from outpatients indeterminate (due to the method of administration) but nevertheless very high; 47 percent for potential respondents from the community and 44 percent from the possible respondents to the postal survey. The latter samples were stratified according to socio-demographic characteristics of the Australian population.

The subsequent process of item reduction used two structural equation programs; for dimensions with fewer than 20 items the EQS program was used. For other dimensions the LISREL program was applied. In each case canonical correlations were used to reflect the ordinal nature of the data from the items. At first the LISREL structural equation program was used to test the hypothesis that all items included in the dimension measured some underlying concept. Secondly the internal structure within each domain was analysed and finally cross loadings between dimensions examined.

Figure 1 shows the final result of this analysis which provides a model that closely fits the data from which it is derived. The instrument consists of 6 dimensions and 20 items. Each of these has between 4 and 6 response levels. Commencing from the right side of Figure 1, the first column of numbers are the gamma coefficients between the dimensions and the overall AQoL latent variable. These are equivalent to standardised correlation coefficients. 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 1 – may also be interpreted as equivalent to correlation coefficients. None is below 0.5 and most above 0.7. Error terms on the individual items in the final right hand column, are generally low for an analysis of individual level data. The CFI of 0.99 is considerably higher than the accepted value of 0.90 and indicates a good fit. The RMSEA of 0.054 is well below 0.08, generally accepted as the minimum level for satisfactory fit. This is an exceptionally good result which underpins the validity and reliability of the model as a representation of the structure underlying the data from our construction sample (Brown and Cudeck 1993,Yu 2002)[5]. The instrument is reproduced in Appendix 1.

[5] Yu, 2002, investigates goodness of fit indices where data deviate substantially from normality and recommends CFI > .95 and RMSEA around .5 top .6 as providing acceptable Type 1 (5) and Type II errors.

Figure 1. Structure of AQoL 2



Notes:

Chi square = 460.73, df=164, P-value=0.0000,RMSEA=0.054, CFI=0.99

(a) (b) (c) From the right, the three sets of numbers represent (a) error terms on each of the items; (b) lambda coefficients (between dimension and items); (c) gamma coefficients (between the AQoL and dimension latent variables).

SOCIAL=Social and family; INDLIV=Independent living; MHEALTH=Mental Health; COPING=Coping; PAIN=Pain; SENSPER=Senses

Table 5 reports the item response level disutilities, calculated from the second postal survey and the transformation to TTO scores described in Appendix 3. The item importance weights, w_i, in Table 6 are calculated from TTO interviews as the item worst score (on a scale from dimension best (0.00) to dimension worst 1.00) health state). These are multiplied by the dimension scaling factor (k_d) which is derived from the item weights and from equation 3 above. The overall or net item weight is used to construct the dimension formulae shown in Box 2.

Table 5. Item Disutilities (TTO Scores) for use in Dimension Models (Mean Values)

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	0.21
4	0.47	1.00	0.83	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	0.34	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			

Notes:
Item best and worst disutilities are set equal to 0.00 and 1.00 respectively. Item best and worst responses are set as endpoints for rating scale evaluations.

Table 6. Item Weights for use in Dimension Models

Dimension \ Item		(-) $k_d\, w_i$	= wt_i	Dimension \ Item		(-) $k_i\, w_i$	= wt_i
Ind Living	1	(0.978) * (.39)	= 0.38	Coping	1	(0.930) * (.42)	= 0.39
	2	(0.978) * (.59)	= 0.59		2	(0.930) * (.64)	= 0.60
	3	(0.978) * (.63)	= 0.63		3	(0.930) * (.77)	= 0.72
	4	(0.978) * (.79)	= 0.79				
Social & Family	1	(0.923) * (.64)	= 0.64	Pain	1	(0.902) * (.63)	= 0. 57
	2	(0.923) * (.70)	= 0.70		2	(0.902) * (.77)	= 0.69
	3	(0.923) * (.51)	= 0.51		3	(0.902) * (.65)	= 0.59
Mental Health	1	(0.983) * (.64)	= 0.64	Sensory	1	(0.851) * (.58)	= 0.49
	2	(0.983) * (.59)	= 0.59		2	(0.851) * (.46)	= 0.39
	3	(0.983) * (.65)	= 0.65		3	(0.851) * (.60)	= 0.51
	4	(0.983) * (.71)	= 0.71				

Similar results are shown for the overall AQoL 2 in Box 2. In this w_d represents the dimension importance weights reported in Table 7 and measured on an AQoL 2 best (DU = 0.00) to death scale (DU = 1.00). The AQoL 2 scaling constant, *k*, is derived from these six weights. The product of the weights and the scaling constant give the effective weights w_d. It is used to derive the overall AQoL 2 formula in Box 2.

Table 7. Dimension Weights for use in AQoL Model

Dimension	(-) k_d x w_d	=	$wt_d^{(i)}$
1. Ind. Living	0.96 x (.49)	=	0.472
2. Social	0.96 x (.47)	=	0.448
3. Mental Health	0.96 x (.50)	=	0.479
4. Coping	0.96 x (.37)	=	0.345
5. Pain	0.96 x (.60)	=	0.592
6. Senses	0.96 x (.65)	=	0.637
AQoL	*W/k* = 1.102/.954	=	1.15

k_d = Dimension scaling constant; w_d = Dimension weight = dimension all worst
wt_d = Final dimension weight (correct for rounded decimal places in col 1, 2)

An example of the use of the formula is given in Box 3.

Box 2. Multiplicative Disutility Equations

Dimensions	
General Formula	$DU_d = \frac{1}{k}\left[1 - \prod_{i=1}^{n}\left(1 - kw_i DU_i\right)\right]; k_d > 0$
Independent Living	$DU_1 = 1.02\left[1 - \left(1 - 0.38du_1\right)\left(1. - 0.58du_2\right)\left(1 - 0.62du_3\right)\left(1 - 0.77du_4\right)\right]$
Social and Family	$DU_2 = 1.08\left[1 - \left(1 - 0.59du_5\right)\left(1. - 0.65du_6\right)\left(1 - 0.47du_7\right)\right]$
Mental Health	$DU_3 = 1.02\left[1 - \left(1 - 0.63du_8\right)\left(1. - 0.58du_9\right)\left(1 - 0.64du_{10}\right)\left(1 - 0.70du_{11}\right)\right]$
Coping	$DU_4 = 1.08\left[1 - \left(1 - 0.39du_{12}\right)\left(1. - 0.60du_{13}\right)\left(1 - 0.72du_{14}\right)\right]$
Pain	$DU_5 = 1.08\left[1 - \left(1 - 0.57du_{15}\right)\left(1 - 0.39du_{16}\right)\left(1 - 0.59du_{17}\right)\right]$
Senses	$DU_6 = 1.18\left[1 - \left(1 - 0.49du_{18}\right)\left(1 - 0.39du_{19}\right)\left(1 - 0.51du_{20}\right)\right]$
General Formula	$DU_{AQoL} = \frac{W}{k}\left[1 - \prod_d \left(\left(1 - kw_d DUx_i\right)\right)\right]; k > 0$
$DU_{AQoL} = 1.150\left[1 - \left(1 - 0.472DU_1\right)\left(1 - 0.42DU_2\right)\left(1 - 0.479DU_3\right)\left(1 - 0.345DU_4\right)\left(1 - 0.592DU_5\right)\left(1 - 0.637DU_6\right)\right]$	

Notes:

W = the conversion factor between the death, full health and multiplicative model (see text)

Box 3. Calculating a utility score: a numerical example

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

Example: Response levels are:

D 1(1,1,2,1); D 2(2,2,3); D 3 (3,2,3,1); D 4(1,1,1); D 5(2,1,1); D 6(2,1,2)

2 Read the 20 disutility scores from Table 3

In the example:

$$D1(0,0.04,0):D2(.07,.19,.65):D3(.39,.14,.33,.00)D4(0,0,0):D5(.13,.0,.0)D6(.03,00,.19)$$

3 Enter the 20 disutility scores into the dimension equations in Box 2

$$DU_1 = 1.02[1-(1-38*0)(1-.58*0)(1-.62*.04)(1-.77*0)] \quad = 0.03$$

$$DU_2 = 1.08[1-(1-.59*.07)(1-.65*.19)(1-.47*.0) \quad = 0.17$$

$$DU_3 = 1.02[1-(1-.63*.39)(1-.66*.14)(1-.64*.33)(1-.7*0) \quad = 0.40$$

$$DU_4 = 1.08[1-(1-.39*0)(1-.60*0)(1-.72*.0) \quad = 0.00$$

$$DU_5 = 1.08[1-(1-.69*.13)(1-.57*0)(1-.57*.0) \quad = 0.10$$

$$DU_6 = 1.18[1-(.1-.4*0.03)(1-.39*0)(1-.51*.19) \quad = 0.12$$

4 Enter the DU_i scores into the AQoL formula Box 1

$$DU_{AQol} = 1.15[1-(1-.472*.03)(1-.448*.17)(1-.479*.4)$$

$$(1-.345*0.0)(1-.592*.1)(1-.637*.12)] = .42$$

5 Convert disutility to utilities from the equation $U = 1 - DU_i$

Dimension Utilities = 0.97; 0.83; 0.6; 1.00; 0.9; 0.88

AQoL U = 0.58

Econometric Modelling

Three sets of multi-attribute health state data were available. As part of the TTO interviews 411 respondents were asked to evaluate 3 or 4 multi-attribute health states selected from 18 holistic ('e'-type) health states which were, in turn, constructed from the AQoL descriptive system. These were designed to include interaction between all combinations of dimensions. From the 365 useable interviews 1042 multi-attribute health state valuations were obtained ('MA.TTOs'). Secondly, the MA.TTO scores were 'deconstructed' to form 'pseudo TTO' health states which describe less severe symptoms and thereby increase the sensitivity of the results to health state values close to full health.[6] Thirdly, a single 'D type health state' value was elicited from each interview respondent for the AQoL all worst health state. All MA (TTO) states were measured on a full health (0.00)-death (1.00) scale.

[6] For example, with 3 dimensions E(U_1 U_2 U_3) PE_1 and PE_2 are created as (U_1 0 U_3) and (0 U_2 0). Scores were assigned by pro rata allocation of the DU of state E between the two states in proportion to the two MA scores for PE_1 and PE_2 derived from the AQoL 2 (multiplicative) model.

The different combinations of these data were employed with the various combinations of the model variables in equation 6. This resulted in 40 basic models: 4 data sets each with 10 combinations of variables which are possible from the variable sets in equation 6. Stage 3 regressions were estimated using a random effects (RE) model[7] and initially evaluated using the conventional Wald statistic.

Table 8 reports the results for three sets of equations estimated with three sets of data. Set A includes the stage 2 estimate of AQoL 2 as the only explanatory variable. Set B also includes dimension variables D_i and dimension interaction terms, D_iD_j. Set C adds the shift dummy variables defined in equation 6. The three equations in each set correspond with the use of observed multi-attribute TTO data only (E); the inclusion of pseudo MA data (P) and the addition of instrument all worst scores, D.

[7] Since respondents were asked to answer between 3 and 6 E type questions an RE model was used to correctly account for clustering in the data.

Table 8. Stage 3 regression results: MA-TTO on AQoL stage 2 predicted values

	Set A			Set B			Set C		
	M1	M2	M3	M4	M5	M6	M7	M8	M9
	E (1)	EP (1)	EPD (1)	E (1)	EP (1)	EPD (1)	E (1)	EP (1)	EPD (1)
n	1042	4136	4501	1042	4136	4501	1042	4136	4501
Wald Chi 2(1)	655	2156	4368	908	2764	5087	1451	6917	8677
AQoL (α)	1.84	1.15	1.51	2.07*	1.14	1.46	1.82	1.46	1.45
D1						0.52*			
D2					0.24*	0.43	-2.54		
D3							-1.97		
D5				-1.5			-1.29*		
D6								0.92	0.70
D_1D_2				-13.4	-3.89*	-3.97	-21.49	-2.84	-4.69
D_1D_5				6.9					
D_1D_6				19.1		8.47			
D_2D_4								2.57*	
D_2D_5								-4.88*	
DD									-1.42**
D_3D_4						2.57*			-2.23*
D_3D_6				-27.5	-13.58	-4.02*	-10.52*	-18.83	
L1									
L2							0.99	0.42	0.42
L3							2.53	1.16	1.10
L4							6.42	2.89	2.68
L5									5.30

(1) All coefficients significant at 0.000 level unless designated
* significant to 0.005 level
** significant at 0.05 level

Four general conclusions may be drawn from Table 8. First, results confirm the expectation that utilities observed in the level 2 multiplicative model inflate DU scores. All models predict disutilities less than predicted by the multiplicative model. Secondly, this effect is not uniform across the dimensions. Thirdly, interaction terms were significant in regressions which included them. Finally, the choice of data set had a significant effect upon results, with the larger data bases producing better fitting models.

Table 9 reports the correlation between the scores, predicted by the 9 models, and the stage 1 AQoL 2 score. Unsurprisingly these are very high and do not provide a basis for discriminating between models.

Table 9. Correlation matrix

	AQoL	M1	M2	M3	M4	M5
AQoL [1]	1.00					
Mean [2]	0.89					
M1	0.98	1.00				
M2	0.99	0.99	1.00			
M3	0.99	0.99	0.99	1.00		
M4	0.94	0.96	0.95	0.96	1.00	
M5	0.99	0.98	0.99	0.98	0.94	1.00
M6	0.98	0.99	0.98	0.99	0.98	0.97
M7	0.92	0.94	0.92	0.93	0.92	0.90
M8	0.93	0.95	0.94	0.95	0.96	0.93
M9	0.95	0.97	0.96	0.97	0.95	0.93
	M6	M7e	M8	M9		
M6	1.00					
M7	0.93	1.00				
M8	0.96	0.95	1.00			
M9	0.96	0.97	0.98	1.00		

[1] AQoL multiplicative model
[2] Mean TTO data
[3] DU estimates using RE regression models

4. Validation and Model Selection

In addition to the Wald Chi 2 statistics, a number of procedures were undertaken to test the goodness of fit of the models. These were (i) the 'internal predictive power'; - a quasi diagnostic test of the explanatory power of stage 3 estimates but, more importantly, a test of bias in the estimate of the TTO scores which is equivalent to a test of the strong interval property (see Footnote 3); (ii) analysis of absolute errors; (iii) comparison of predicted and actual change between health states; and (iv) external validation – comparison of actual and predicted scores from an independently collected dataset.

Internal Predictive Power

To conduct this test the predicted TTO score from the stage 3 models were used as the (only) independent variable in a linear regression to explain MA (TTO) scores. Since RE modelling should result in an unbiased estimate of the dependent variable, OLS modelling should be used to test for bias. For each model the constant term was initially suppressed. An unbiased model would result in b = 1.00 in equation 7 below:

In the second set of regressions a constant term was included to test the (more demanding) null hypothesis: that a = 0.00; b = 1.00 in equation (8) below

$$MA\ (TTO) = b.M_i + e \qquad ...(7)$$

$$MA\ (TTO) = a + bM_i + e \qquad ...(8)$$

Where M_i = stage 3 utility predicted by model M_i

The two regressions test whether or not models provide an unbiased estimate of average and marginal TTO scores respectively. The second of these is particularly relevant for the estimation of marginal QoL changes effected by a health program.

Tables 10.1 to 10.4 present results using individual (E + D) data and mean (E) data. Coefficients for 'b' in Table 10.1 indicate a close correspondence between average predicted and average observed data except for models 4 and 8. However results in Table 10.2 indicate that slope coefficients are less than unity and suggest that stage 3 coefficients may underestimate the full effect of incremental change. Models 4 and 7 perform particularly badly on this test with model 9 having the least bias. Using mean data (Tables 10.3, 10.4) model 7 and model 8 obtain low R^2 coefficients; models 1 to model 3 have relatively small b coefficients in the range 0.69-0.73, but high R^2 coefficients in the range 0.90-0.95 indicating high explanatory power. Once again M9 provided the best estimate, explaining 90 percent of variation and with a 'b' coefficient within 6 percentage points of the gold standard value of 1.00.

Table 10. Stage 4 OLS regressions: Observed (MA) TTO on predicted TTO, Models 1-9

10.1	$MA(TTO) = bM_i + e$			Individual E + D Data, n = 1403					
	M1	M2	M3	M4	M5	M6	M7	M8	M9
B	0.84	0.86	0.86	0.79	0.87	0.86	1.07	1.32	1.07
R^2	(0.88)	(0.88)	(0.88)	(0.50)	(0.71)	(0.86)	(0.38)	(0.49)	(0.86)
10.2	$MA(TTO) = a + bM_i + e$			Individual E + D Data, n = 1407					
	M1	M2	M3	M4	M5	M6	M7	M8	M9
A	0.08	0.02	0.02	0.34	0.17	0.04	0.40	0.33	0.11
B	0.76	0.84	0.84	0.42	0.68	0.80	0.40	0.64	0.86
R^2	(0.75)	(0.74)	(0.74)	(0.18)	(0.41)	(0.68)	(0.07)	(0.12)	(0.72)
10.3	$MA(TTO) = a + bM_i + e$			Mean E + D Data, n = 74					
	M1	M2	M3	M4	M5	M6	M7	M8	M9
A	0.06	0.04	0.04	0.07	0.03	0.03	0.04	0.02	0.03
B	0.74	0.73	0.73	0.74	0.81	0.75	1.13	1.27	1.05
R^2	0.95	0.93	0.93	0.62	0.84	0.92	0.40	0.52	0.90
10.4	$MA(TTO) = a + bM_i + e$			Mean E Data, n = 19					
	M1	M2	M3	M4	M5	M6	M7	M8	M9
A	0.10	0.05	0.05	0.15	0.02	0.04	0.17	0.13	0.10
B	0.69	0.73	0.73	0.62	0.84	0.76	0.88	1.01	0.96
R^2	0.93	0.90	0.90	0.40	0.72	0.88	0.19	0.22	0.92

Notes:

The R^2 test statistic for regressions omitting the constant cannot be interpreted in the normal way (as the percentage of variation 'explained'). Nevertheless they reflect a satisfactory fit except for models 7 and 8.

The prediction of incremental change was tested next and results are reported in Table 11. For each of the four best fitting models the difference between various combinations of observed TTO scores were subtracted from the difference predicted by model scores, that is, observed (TTO_i - TTO_j) was subtracted from predicted ($AQoL_i$ - $AQoL_j$). The reported result is therefore the absolute error from model prediction as compared with gold standard TTO scores[8].

[8] Because different combinations of health states were administered to different respondents there were relatively few health states where a comparison of this form could be made.

Table 11. Errors in estimates of incremental changes [TTO(ei)-TTO(ej)]-[Model ei – model ej] Absolute values

variable	Model 1 OBS-Pred	Model 3 OBS-Pred	Model 6 OBS-Pred	Model 9 OBS-Pred
column	2	3	4	5
chg e1 to e2	0.00	0.01	0.01	0.01
chg e1 to e3	0.04	0.04	005	0.05
chg e2 to e3	0.03	0.05	0.03	0.03
chg e4 to e5	0.23	0.25	0.29	0.11
chg e4 to e6	0.19	0.16	0.10	0.06
chg e5 to e6	0.42	0.42	0.39	0.18
chg e7 to e8	0.03	0.03	0.00	0.02
chg e7 to e9	0.16	0.20	0.15	0.04
chg e8 to e9	0.13	0.17	0.15	0.01
chg e10 to e11	0.32	0.36	0.33	0.01
chg e10 to e12	0.26	0.24	0.16	0.04
chg e11 to e12	0.06	0.11	0.17	0.03
chg e13 to e14	0.04	0.04	0.03	0.02
chg e13 to e15	0.00	0.00	0.02	0.14
chg e14 to e15	0.04	0.04	0.01	0.11
chg e16 to e17	0.09	0.07	0.16	0.08
chg e16 to e18	0.00	0.01	0.04	0.16
chg e17 to e18	0.09	0.09	0.12	0.08
Average*	0.12	0.13	0.12	0.07
Error 0.00-0.049	8	7	7	9
0.05-0.099	3	3	1	4
0.1-0.39	44	7	10	5
0.4*	3	1	0	0
Total				

*Average of absolute value ignoring the sign

Results do not clearly indicate the relative goodness of fit of the five models. M9 resulted in the smallest number of relatively large errors and the lowest average error. However the chief conclusion to be drawn from the table is that the pattern of errors created by the different models differs and some of the real incremental changes are very different from estimates. This suggests that the database may have underrepresented these health states in the interview protocol or that co linearity may have resulted in selected error. It is difficult to draw stronger conclusions from this analysis.

External Validation: The VisQoL Study

An independent study was recently completed to recalibrate AQoL 2 for the visually impaired (Misajon et al. 2005). From this 752 MA-TTO (VisQoL (TTO)) values were obtained for the visually impaired and the general population using 29 health states from the AQoL 2 descriptive system. Individual and mean data were used to estimate equation 9.

$$\text{VisQoL (TTO)} = a + b\ \text{AQoL} \qquad \ldots (9)$$

As previously, an unbiased model would result in coefficients of a = 0.00; and b = 1.00. Results are reported in Table 12 below. They indicate poor predictive power at the individual level. Unreported R^2 values were also small. This result is to be expected. Most of the variation in the dependent variable is between individuals with the same health state and there is only a single predicted AQoL score to 'explain' this variation; that is, most of the variation cannot be explained because of the construction of the test. The use of mean data overcomes this problem as it is concerned with 'between' health state variation. It is also the more relevant test as CUA studies require mean, not individual, scores. Table 12 row 3 indicates very high explanatory power of all models with between 63 and 79 percent of variation explained. From row 2, coefficients for models M1 to model M6 indicate that VisQoL (TTO) scores vary by only 52-59 percent as much as predicted AQoL model scores over the range of observations. However model 7 to model 9 provide estimates of b which are very close to unity and imply little bias. Results from the previous five tables are summarised in Table 13. From these two tables model 9 is most consistently the best performing model and is therefore the one recommended for use.

Table 12. Results from the VisQoL validation study

Coefficient	**Row**	**Stage 4 –type regressions: VisQoL (TTO) = a + b M_i + e**								
(data)		M1	M2	M3	M4	M5	M6	M7	M8	M9
b (individual) n=752		0.43	0.45	0.43	0.41	0.44	0.46	0.76	0.64	0.65
b (mean) n=	1	0.59	0.54	0.56	0.55	0.52	0.58	1.08	0.86	0.87
R^2 (mean)	2	0.79	0.78	0.79	0.78	0.79	0.78	0.63	0.73	0.74

Notes:
'a' coefficients, R^2 (individual data) not reported
Relatively good results
Relatively poor results

Table 13. Summary of selection criteria

	M1	M2	M3	M4	M5	M6	M7	M8	M9
Stage 3	Various databases								
Wald	655	2156	4368	908	2764	5687	1431	6917	8677
Stage 4a	n = 1407								
B	0.84	0.86	0.86	0.79	0.87	0.86	1.07	1.32	1.02
R^2	0.88	0.88	0.88	0.50	0.71	0.86	0.38	0.49	0.86
Stage 4b	n = 74								
B	0.74	0.73	0.72	0.74	0.81	0.75	1.13	127	1.05
R^2	0.95	0.93	0.93	0.62	0.84	0.92	0.40	0.52	0.90
Error, predicted change, n = 19									
	0.12		0.13			0.13			

Notes:

Relatively good results
Relatively poor results

VisQoL see Table 10

5. Discussion and Conclusion

AQoL 2 is a more complex MAU instrument than others reported in the literature. This was justified by the need for a MAU instrument to evaluate public health in addition to acute health treatments and therefore the need to increase the sensitivity to interventions where the benefits are associated with the change in the level of handicap and where 'within the skin' description is less likely to provide an indication of utility. The applicability of different instruments will vary with the type of intervention and its chief effects.

The greater complexity of the instrument brought its own problems in the construction methods. In particular, it was difficult to envisage a descriptive system of this complexity which achieves construct validity without the use of the appropriate psychometric construction methods. With more elements and dimensions potentially subtracting from utility constructural and preference dependence become of greater concern.

While administration of the instrument was relatively simple – typically taking 8 to 10 minutes – calculation of utilities was complex (although the task provided 6 dimension scores as a bonus!) Copies of the AQoL 2, user registration forms, and the scoring algorithm are available from the CHE Monash website (http://www.buseco.monash.edu.au/centres/che), the Melbourne WHOQOL website (http://www.psychiatry.unimelb.edu.au/qol/) or from the authors.
The AQoL project also sought to contribute to the methodology of MA instrument design and construction. Its use of psychometric methods appears to be unique. Its multi level structure and utility modelling are innovative. Experimentation suggested that focusing effects probably do occur (Appendix 2), utility scores do fall somewhat with deliberation (Appendix 5). Negative TTO scores are probably not a serious problem (Appendix 3). Perhaps of greater importance the project has introduced validation tests which move beyond correlation and attempt to demonstrate the existence of a strong interval property in the final instrument scores.

The resulting AQoL 2 model strongly suggests that the structure of preferences for health states does not lend itself to simple modelling. This arises, fundamentally, from the complexity of the underlying construct. Results here suggest both structural and preference dependency in the simple modelling. The methods developed in this paper should, in principle, minimise both of these sources of error. Results provide confidence in this conclusion. It is difficult, however, to put the outcome of these tests in broad perspective as there are no comparable test results in the literature. In absolute terms however they suggest that AQoL 2 is a good instrument.

References

Bowling, A,. 2001. Measuring Disease: A Review of Disease-Specific Quality of Life Measurement Scales. Second edition. Open University Press: Buckingham.

Bowling, A., 2005. Measuring Health: A Review of Quality of Life Measurement Scales. Third edition. Open University Press: Maidenhead.

Brazier, J.E., Harper, R., Thomas, K., Jones, N., Underwood, T., 1998. Deriving a preference based single index measure from the SF36. Journal of Clinical Epidemiology 51, 1115-1129.

Brazier, J.E., Roberts, J., Deverill, M., 1999. The Estimation of a Utility Based Algorithm from the SF-36 Health Survey, Report Prepared for Glaxo Welcome, Mimeo.

Brazier, J.E., Roberts, J., Deverill, M., 2002. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 211, 271-92.

Brown, M.W., Cudeck, R., 1993. Alternate ways of assessing model fit. In K.A. Bollen, J.S. Long (Eds), Testing Structural Equation Models, Sage: Newbury Park.

Day, N., Richardson, J., Peacock, S., Hawthorne, G., Iezzi, A., 2007. Development of the Structure of the Assessment of Quality of Life (AQoL) 2 Health Related Quality of Life Instrument, Research Paper, Centre for Health Economics, Monash University: Melbourne (forthcoming).

Dolan, P., Gudex, C., Kind, P., Williams, A., 1995. A social tariff for EuroQoL: Results from a UK General Population Survey, Discussion Paper No 138, Centre for Health Economics, University of York: York.

EuroQoL Group: Rabin, R., De Charro, F., 2001. EQ-5D: A measure of health status from the EuroQoL Group. Annals of Medicine 33, 337-343.

Feeny, D., Torrance, G., et al., 1996. Health Utilities Index. In B Spilker (Ed) Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition, Lippincott-Raven Publishers: Philadelphia 239-252.

Feeny, D., Furlong, W., et al., 2002. Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. Medical Care 40, 113-128.

Furlong, W.J., Feeny, D.H., Torrance, G.W., Barr, R.D., 2001. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. Annals of Medicine 33, 375-384.

Hawthorne, G., Richardson, J., 1995. An Australian MAU/QALY Instrument: Rationale and Preliminary Results, Working Paper 49, Centre for Health Program Evaluation, Monash University: Melbourne.

Hawthorne, G., Richardson, J., et al., 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: 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., Richardson, J., Day, N.A., 2001. A comparison of the assessment of quality of life (AQoL) with four other generic utility instruments. Annals of Medicine 33, 358-370.

Horseman, J., Furlong, W., Feeny, D., Torrance, G., 2003. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health and Quality of Life Outcomes 1, 54 (http://www.hqlo.com/content/1/1/54).

Kaplan, R.M., Ganiats, T., et al., 1996a. The quality of wellbeing scale. Medical Outcome Trust Bulletin 4, 2-3.

Kaplan, R.M., Anderson, J.P., 1996b. The general health policy model: an integrated approach. In B Spilker (Ed) Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition, Lippincott-Raven Publishers: Philadelphia 309-22.

Kind, P., 1996. The EuroQoL instrument: an index of health related quality of life. In B Spilker (Ed) Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition, Lippincott-Raven Publishers: Philadelphia.

Misajon, R., Hawthorne, G., Richardson, J., Barton, J., Peacock, S., Iezzi, A., Keeffe, J., 2005. Vision and quality of life: the development of a utility measure. Investigative Ophthalmology & Visual Science 46, 4007-4015.

Peacock, S., Richardson, J., Hawthorne, G., Day, N.A., Iezzi, A., 2004. 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: Melbourne.

Richardson, J., 1994. Cost utility analysis: what should be measured. Social Science and Medicine 39, 7-21.

Richardson, J., Hawthorne, G., 1999. Negative Utility Scores and Evaluating the AQoL All Worst Health State, Working Paper 73, Centre for Health Program Evaluation, Monash University: Melbourne.

Richardson, J., Hawthorne, G., Day, N.A, Peacock, S., Iezzi, A., 2004a. 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: Melbourne.

Richardson, J., Peacock, S., Hawthorne, G., Day, N.A., Iezzi, A., 2004b. 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: Melbourne.

Robinson, A., Spencer, A., 2006. Exploring challenges to TTO utilities: valuing states worse than dead. Health Economics 15, 393-402.

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

Sintonen, H., 2001. The 15D instrument of health-related quality of life: properties and applications. Annals of Medicine 33, 328-336.

Ubel, P., Lowenstein, G., et al., 2001. Do non-patients underestimate the quality of life associated with chronic health conditions because of a focusing illusion? Medical Decision Making 21, 190-199.

von Winterfeldt, E., Edwards, W., 1986. Decision Analysis and Behavioural Research. Cambridge University Press: Cambridge.

Ware, J., Snow, K., Kosinski, M., Gandek, B., 1993. SF-36 Health survey: manual and interpretation guide. The Health Institute, New England Medical Centre: Boston.

WHOQOL Group, 1998. Development of the World Health Organization (WHOQOL-BREF quality of life assessment. Psychological Medicine 28, 551-8.

World Health Organization (WHO), 1980. International Classification of Impairments, Disabilities and Handicaps, WHO: Geneva.

Yu, C.Y., 2002. Evaluating Cut-off Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes. Dissertation for PhD. University of California: Los Angeles.

Appendix 1. AQoL 2 Questionnaire

Assessment of Quality of Life (AQoL) Mark 2

Dimension 1: Independent Living

Q1 How much help do I need with household tasks (eg 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.

Q2 Thinking about how easy or difficult it is for me to get around by myself outside my house (eg 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.

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.

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.

Dimension 2: Social and Family

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

. very happy
. generally happy
. neither happy nor unhappy
. generally unhappy
. very unhappy

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.

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.

Dimension 3: Mental Health

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

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

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

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

Q10 How often do I feel sad?

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

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.

Dimension 4: Coping

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.

Q13 How often do I feel in control of my life?

. always
. mostly
. sometimes
. only occasionally
. never.

Q14 How much do I feel I can cope with life's problems?

. completely
. mostly
. partly
. very little
. not at all.

Dimension 5: Pain

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.

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.

Q17 How often does pain interfere with my usual activities?

. never
. rarely
. sometimes
. often
. always

Dimension 6: Senses

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.

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.

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

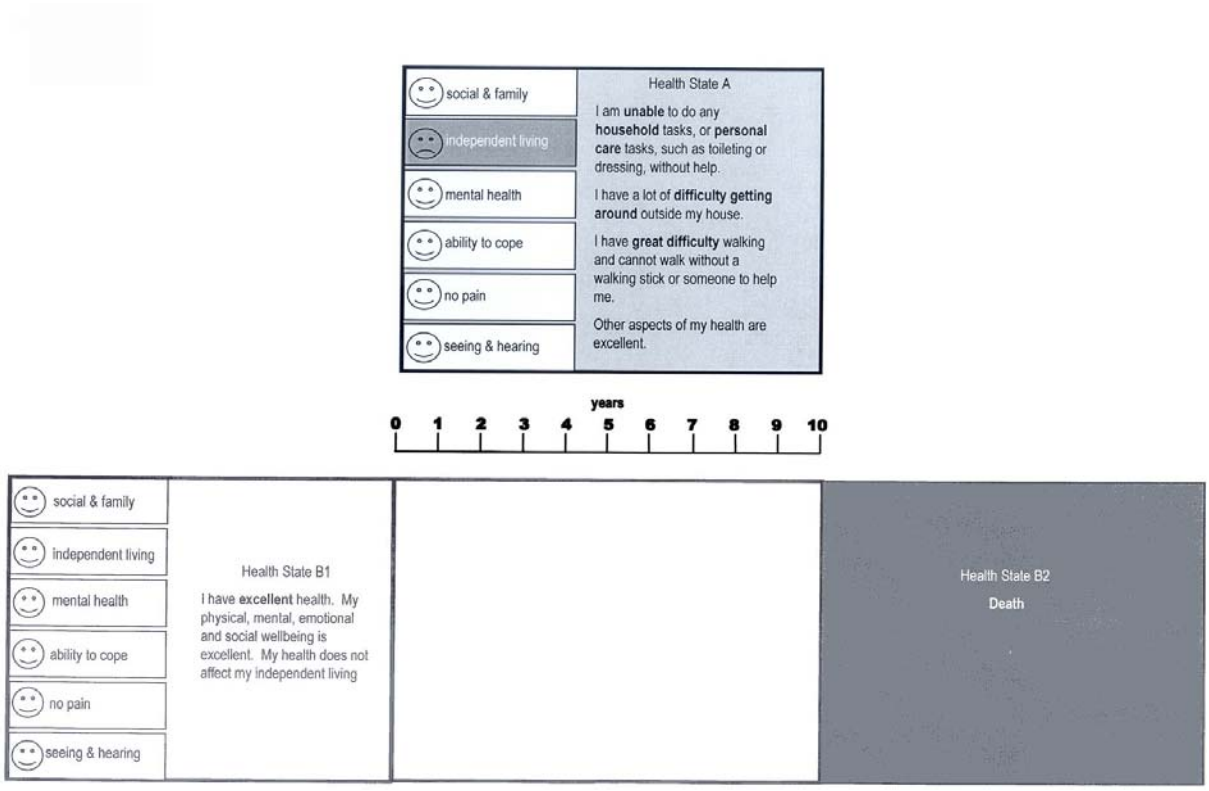
. I have no trouble speaking to them or understanding what they are saying
. I have some difficulty being understood by people who do not know me. I have no trouble understanding what others are saying to me.
. I am understood only by people who know me well. I have great trouble understanding what others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

Appendix 2. Artwork

Figure A1 represents the artwork and slide board used in the TTO elicitation for AQoL 2. Its distinctive feature is the depiction of all dimensions in both words and pictures. At the commencement of the interview each dimension was fully described. Interviewees were reminded of the good health dimensions at the commencement of each trade-off.

TTO disutility results from this procedure were lower than equivalent disutility scores obtained from the AQoL 1 when good health dimensions were referred to but not depicted.

Figure A1. Artwork



Appendix 3. Negative utilities

For a health state worse than death, x years with value V plus (10-x) of good health, U = 1 are equal to death, U = 0.

$$x.V + (10-x).1.00 = 0$$

$$V = -(10-x)/x$$

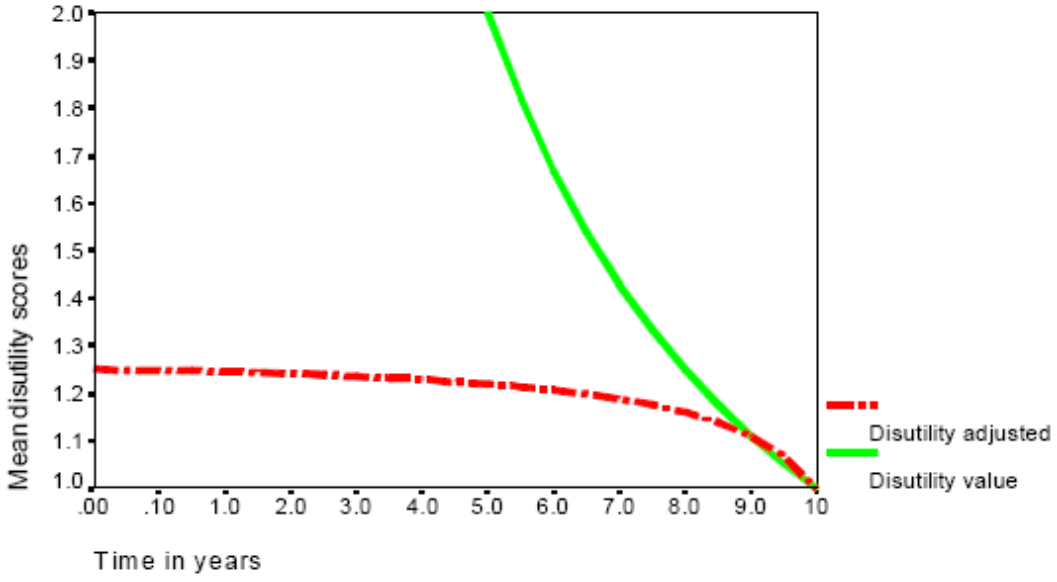
As x approaches zero V approaches minus infinity.

There are an infinite number of transformation functions which constrain negative disutilities between zero and any nominated maximum value. One such function is a modification of equation 2 where c is the lowest permitted value and n determines the transformation function for the negative score.

$$DUA = (1 + c) + 1/(nV - 1/c)$$

There is no logic in setting the maximum DU equal to -1 as selected in the EQ-5D. The apparent symmetry with positive utility is illusory in logic and has no empirical basis (Richardson and Hawthorne 2001). Based upon a survey of 116 community members a median value of 10 was obtained for death on a worst-best imaginable health state rating scale. For this, plus a number of other reasons a maximum possible negative score of -0.25 was selected and a value of n equal to 28.6. This collapses the transformation into the equation DU = 1.25 + 1/(nV-4). This is plotted against untransformed disutility, V, in Figure A2.

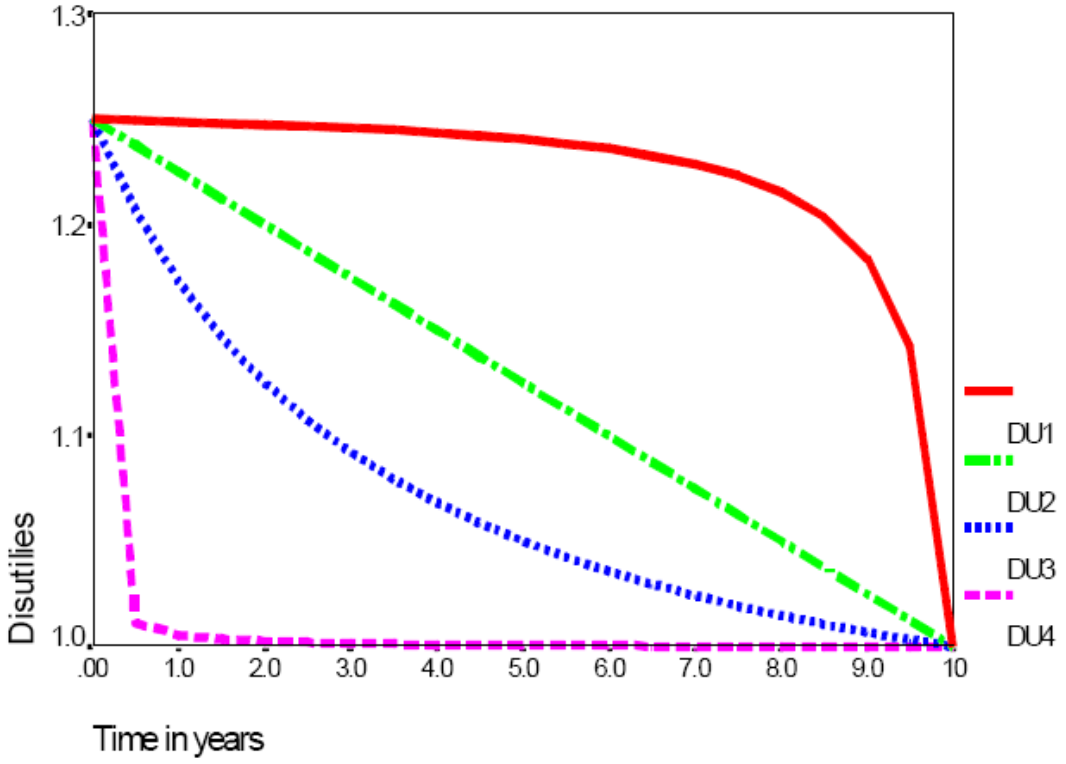
Figure A2. Unadjusted vs adjusted disutilities



Source: Richardson and Hawthorne (1999 p 9)

The effect of varying n is shown in Figure A3. The relationship between this function and that selected by the EQ-5D is shown in Figure A4.

Figure A3. Four transformation patterns for disutilities



Notes:

Calculated from: $$DUA = c + \frac{1}{nV - \frac{1}{m}}$$

DU1: n = 100

DU2: n = 4

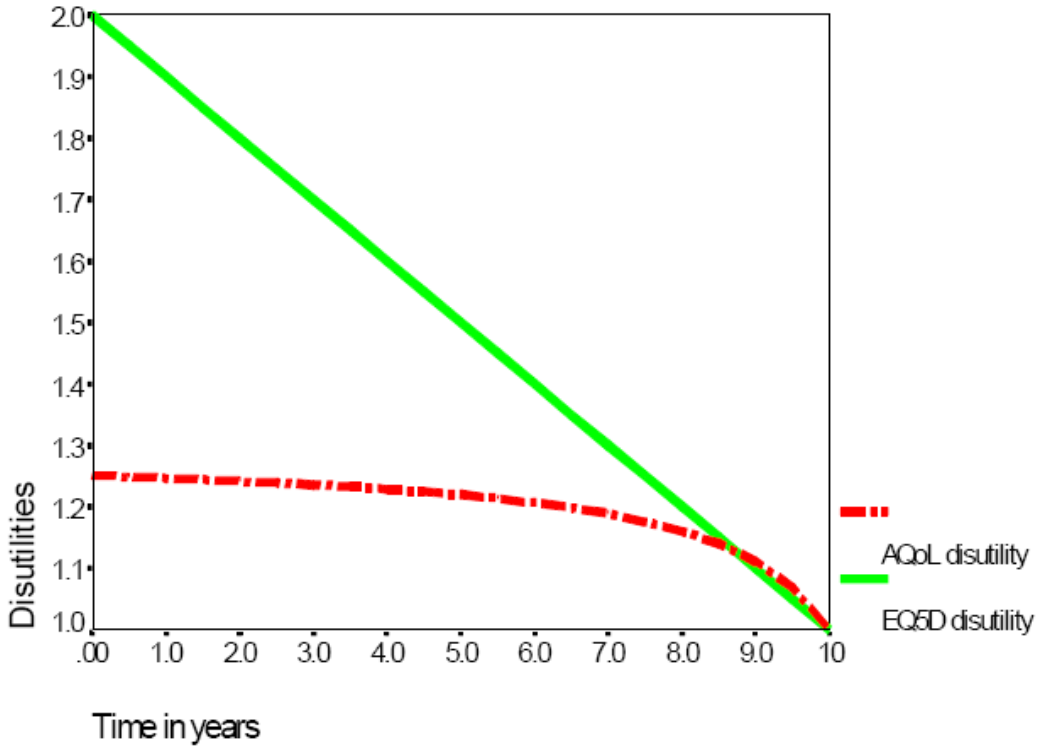
DU3: n = 1

DU4: n = 0.01

Source: Richardson and Hawthorne (1999, p10)

Varying the value of c from -0.2 to -0.5 reduced the value of the AQoL all worst from -0.028 to -0.205 or by 17 percent. From the TTO interviews 76 percent had negative scores for the AQoL all worst. The alteration in the transformation function would therefore alter utility scores by 13 percent. This implies that with utilities in the range 0.7-1.0 reducing the maximum disutility from -0.2 to -0.5 would affect overall utility scores by less than 4 percent.

Figure A4. AQoL vs EQ-5D disutility paths



Source: Richardson and Hawthorne (1999 p11)

Appendix 4. Rating scale – TTO Transformation

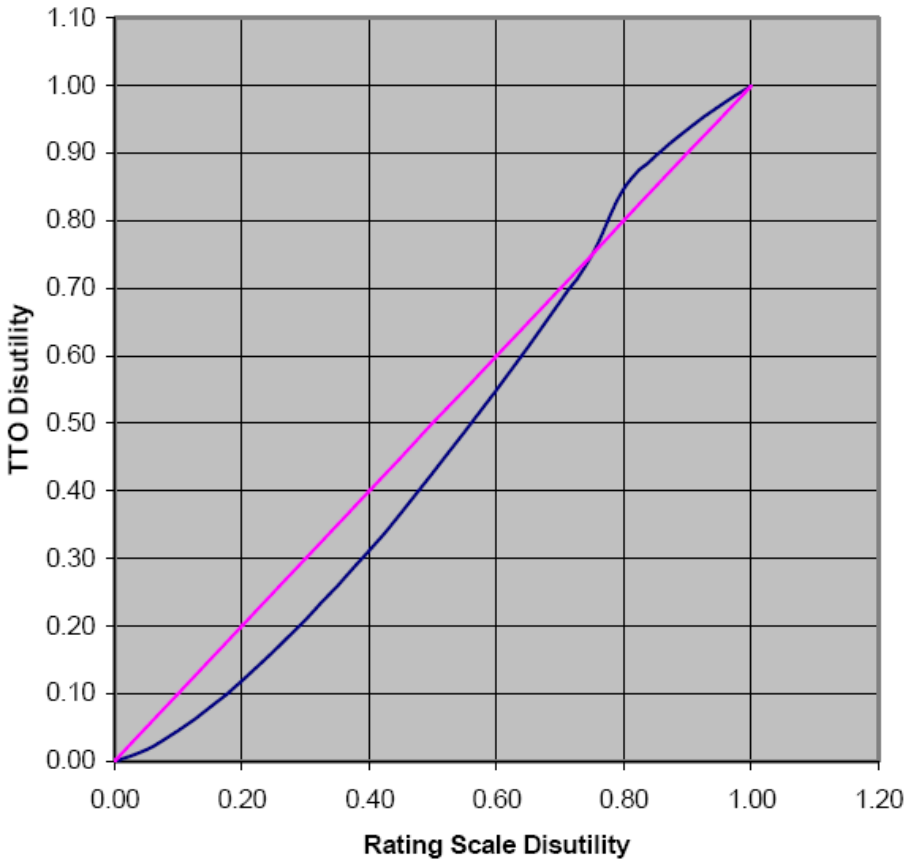
Item response scores were obtained from a postal survey plus a rating scale-TTO transformation based upon a sub-set of respondents providing item worst data on both scales. Relative to TTO scores the RS constant traits observations in the middle of the scale. This results in an 'S' shaped relationship around the 45^{o} line when TTO is on the vertical and rating scale on the horizontal axes. To capture the concavity and convexity above and below the cross over point which occurred, empirically, at (0.75, 0.75) the two functions below were used:

$TTO_1 = .75(1.33\ RS)^{1.394}$ ..., RS< 0.75

$TTO_2 = .75 + .25[(RS-1.4).4]^{.588}$ ..., RS > 0.75

The resulting transformation is plotted in Figure A5.

Figure A5. Two part power function transformation of RS into TTO disutilities.



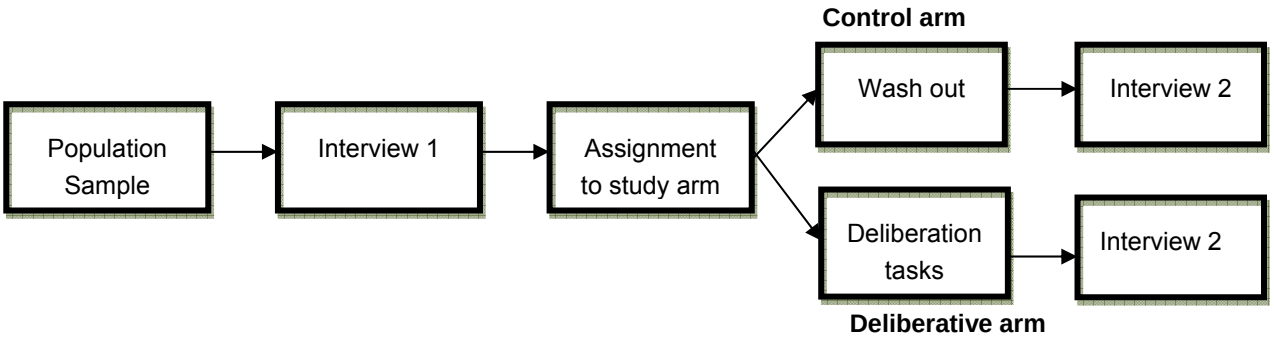
Notes:

Power function point of inflexion at (0.75, 0.75) 45 degree line shows where TTO Disutility = RS Disutility

Appendix 5. The Effect of Deliberation

TTO disutilities were obtained twice, separated by an interval of at least two weeks. Respondents were assigned to a control or 'intervention' arm where the latter took the form of a set of deliberative tasks to be completed at home. These included discussion of questions, answers and reasons with family members and friends. Respondents were asked to speak with the person they might discuss health related problems with in real life, eg spouse, close friend or relative. The design of the experiment is shown in Figure A6 below.

Figure A6. AQoL 2 Deliberative Design



Source: Peacock, Richardson et al. (2004)

Results were tested for self-selection and over-sampling and differences tested using paired t-test (two tier and Pearson chi square test).

Results are presented and analysed in Peacock et al. (2004). These demonstrated no effect attributable to the deliberation task but an overall decline in disutility scores of about 4 percent in both groups. This must be attributed to the effect of familiarity with the task in this second interview and a lessening of the possible 'shock-horror' effect of living in a poorer health state. Both these effects indicate a form of 'deliberation' or 'acceptance/adaptation' which implies some exaggeration in the value of spontaneous disutility scores. The magnitude of the distortion is, however, small. AQoL 2 used the second set of disutility scores.