CENTRE FOR HEALTH PROGRAM EVALUATION

WORKING PAPER 140

A Comparison of Five Multi Attribute Utility Instruments

Paper presented to the Twenty Second Australian Conference of Health Economists 2000

Associate Professor Graeme Hawthorne
Principal Research Fellow
Australian Centre for Posttraumatic Stress Disorder
The University of Melbourne

Professor Jeff Richardson
Professor and Director
Health Economics Unit, Centre for Health Program Evaluation
Monash University

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

June, 2003
ISSN 1325-0663

CENTRE PROFILE

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

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

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

PUBLICATIONS

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

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

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

or

By downloading from our website
Web Address http://chpe.buseco.monash.edu.au

ACKNOWLEDGMENTS

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

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

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

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

Table of Contents

List of Tables

List of Figures

Abstract

This paper presents the results of the validation study carried out to evaluate the Assessment of Quality of Life (AQoL) Instrument for the measurement of health related quality of life and utility. It involves, inter alia, the largest comparison of utility instruments that has been carried out to date. The five instruments included in the study are the AQoL, the Canadian HUI III, the Finnish 15D, the EuroQoL (EQ5D) and the SF36 with UK utility weights as quantified by Brazier (1998). The paper compares: (i) the absolute utility score obtained by different sub-populations; (ii) instrument sensitivity; (iii) the incremental differences in utility between different health states; (iv) the structural properties of descriptive systems; and (v) a limited comparison with a Time Trade-Off (TTO) assessment of own health by individuals. Using these criteria the AQoL performs very well. Its predicted utilities are very similar to those obtained from the HUI. There is evidence that the AQoL has greater sensitivity to health states than other instruments and its psychometric properties, as usually judged, are excellent. Despite this, it is concluded that, at present, no single MAU system can claim to be the gold standard and that researchers should select an instrument that is sensitive to the health states which they are investigating and that caution should be exercised in treating any of the instrument results as representing a utility score which truly represents a trade-off between life and health related quality of life.

A Comparison of Five Multi Attribute Utility Instruments

1 Use and Abuse of MAU Instruments[1]

The quantification of 'utility' in cost utility analysis (CUA) requires two broad tasks. First, the health state under investigation must be described; secondly, a scaling technique such as the time trade-off (TTO) or standard gamble (SG) must be used to attach a numerical value to the health state. This value should measure the strength of a person's preference (utility) for the health state. Two broad approaches to this two stage procedure have normally been used[2], namely, holistic (or 'composite') and multi-attribute utility (MAU) measurement (Torrance 1986). With the first of these, a scenario or vignette is constructed which describes the health state (Step 1). The entire scenario is then 'scaled' (Step 2): ie a survey is conducted specifically to elicit 'utility' values for the scenario. With the second approach a generic 'descriptive system' or 'descriptive instrument' is created which is capable of describing a wide range of health states and utility weights are attached to every possible state. This is normally done by measuring a limited number of 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'[3]. The model may either be derived by econometric analysis of the observed utilities (as with the EuroQoL (Williams 1995)) or through the use of decision analytic techniques to fit the simple additive model (as used in the Quality of Wellbeing Instrument (QWB) (Kaplan et al 1996) and 15D (Sintonen and Pekurinen 1993)) or a multiplicative model (the Health Utilities Index (HUI I, II and III) (Feeney, Torrance et al 1996). The fully scaled MAU instrument may then be used to estimate the utility of all possible health states described by the models' descriptive system.

Both approaches have strengths and weaknesses. Holistic measurement permits the use of a description or vignette which is tailored to a particular health state. This may include unique aspects of the health state, its content, its consequences, the process of health delivery risk or prognosis. Validation of health state specific vignettes, however, is seldom, if ever, carried out. By contrast, the generic 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 of the health state or intervention. However, this approach may, in principle, be based upon a descriptive system, the reliability and validity of which can be investigated using standard procedures[4]. After construction, the use of an MAU instrument is cheap 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 Section 1 is an edited version of Richardson and Hawthorne (1999) Section 1.

2 In principle, these two steps can be collapsed by asking patients directly the value of the health state that they are currently experiencing. In practice this approach has seldom been used. Some results from the use of this technique are reported below.

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

4 Essentially, HRQoL is a psychometric concept, as are utilities. They cannot be directly measured, but are uniquely individual. Although instruments can be developed from other measurement – traditions such as clinometrics, economics or decision-making – this property of HRQoL suggests that the application of selected methods and procedures drawn from psychometrics will be particularly appropriate during instrument construction.

To date, only a handful of generic instruments have attempted to measure utility; *viz.* the UK Rosser-Kind Index (Rosser 1993), the US QWB (Kaplan, Ganiats et al 1996), the Canadian HUI instruments (Feeney, Torrance et al 1996), the Finnish 15D (Sintonen and Pekurinen 1993) and the European EuroQoL (Kind 1996). More recently the WHO has constructed the WHOQoL (to date without utility weights). Brazier has provided a utility scoring algorithm for the SF36 (Brazier 1998); and, finally, working for the World Bank and the WHO, Murray and Lopez (1996) have published 'disutility' weights for the different health states required for the construction of Disability Adjusted Life Years (DALYs) and used these to quantify the burden of every disease in every country in the world. The Assessment of Quality of Life (AQoL) developed by the present authors is the most recently developed of the MAU instruments. Some of the characteristics of these instruments are summarised in Table 1.

Table 1 Major MAU Scales

Scale	Coverage (a)	Type of description (b)	N. dimensions	Valuing method (c)	Psychometric properties (d)		Combination model	Instrument boundaries (e)
					Construct	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/ disability	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	8	VAS/SG	No	Yes	Multiplicative	–0.36 — 1.00
EQ5D	X	Impairment/ disability	5	TTO	No	No	Regression/ Additive	–0.59 — 1.00
DALY	XX	Disease	N/A	PTO	No	No	RS/PTO (f)	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:

a = Coverage of the HRQoL universe, as defined by a review of 14 HRQoL instruments, 1971–1993 (24). Coding scheme: XX = very poor, X = poor, = good, = very good.

b = Based on WHO classification of diseases and impairments (25).

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

d = Descriptive system constructed following standard psychometric rules for instrument construction (26, 27).

e = 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'.

f = Rating scale validated using the PTO

The present paper summarises results from a large scale survey designed specifically to validate – test – the AQoL through its comparison with 4 other widely used instruments. To our knowledge it is the largest such comparative study of utility instruments undertaken to date. In the following sections we briefly describe the survey (Section 3) and present results with respect to average utility scores, the correlation and linear association between instruments (Section 4.1); Instrument sensitivity is then compared (Section 4.2); the internal structure of the different instruments and their relationship to a single coherent concept of HRQoL is then examined (Section 4.3). Finally some limited evidence is presented which compares instrument values for individuals with their self rated time trade-off – time they would sacrifice to move from their current health state to normal health (Section 4.4). It is concluded in Section 5 that none of the instruments at present can claim the status of gold standard.

One of the consequences of the validation study (not reported here) was the finding that one of the five dimensions (illness) of the original AQoL was redundant and its inclusion in the instrument invalidated utility scores. Results in this paper refer to the truncated 12 item 4 dimensional AQoL instrument produced by the deletion of the illness dimension.

2 The Assessment of Quality of Life (AQOL) Instrument

While each of the four other instruments included in this study for comparison with the AQoL has particular strengths, to our knowledge none were constructed using normal psychometric principles to ensure construct validity and structural independence. Consider, for example, this second issue. MAU theory postulates there should be no 'redundancy' amongst items in a descriptive system. That is, a single attribute should not be described in more than one way (von Winterfeldt and Edwards 1986). If redundancy occurs then the (dis)utility of the attribute will be double counted. A sufficient (but not necessary) condition for non-redundancy is that the different scales within the instrument are orthogonal.[5] However, the requirement of non redundancy appears to be in conflict with the need for 'sensitivity' and several instruments have reduced redundancy by the adoption of very simple descriptive systems; but this simplicity has been achieved at the expense of sensitivity.

Other problems also exist. MAU theory specifies that the model used to create an instrument should be determined, inter alia, by the need to achieve ‘preference independence’: the utility – preference score – for an item or dimension should be independent of the item level in other dimensions[6]. Because of the difficulty in testing for preference independence this property, when explicitly considered, is usually assumed to exist or (as with the AQoL) subject to ad hoc and post hoc testing.

Several of the models are additive: it is assumed that scores on one item or dimension are unaffected by scores on other items or dimensions. In order to impose this simple model, item or dimension weights must sum to unity. This implies that the greater the importance of one dimension the lesser importance can be assigned to all other dimensions. This, in turn, imposes

[5] It is not strictly necessary as scales may be 'environmentally correlated', which does not necessarily indicate double counting. Von Winterfeldt & Edwards (1986) illustrate this in the case of a manufacturing plant, the management of which is concerned with the costs of production and distribution. These costs will correlate because each correlates with the scale of production. Despite this, there is no redundancy and each attribute is independently important. Even with this example, however, careful construction of the instrument can eliminate the correlation. There is no necessary reason why scale of production, *unit* production costs and *unit* distribution costs will correlate (if there are no economies of scale).

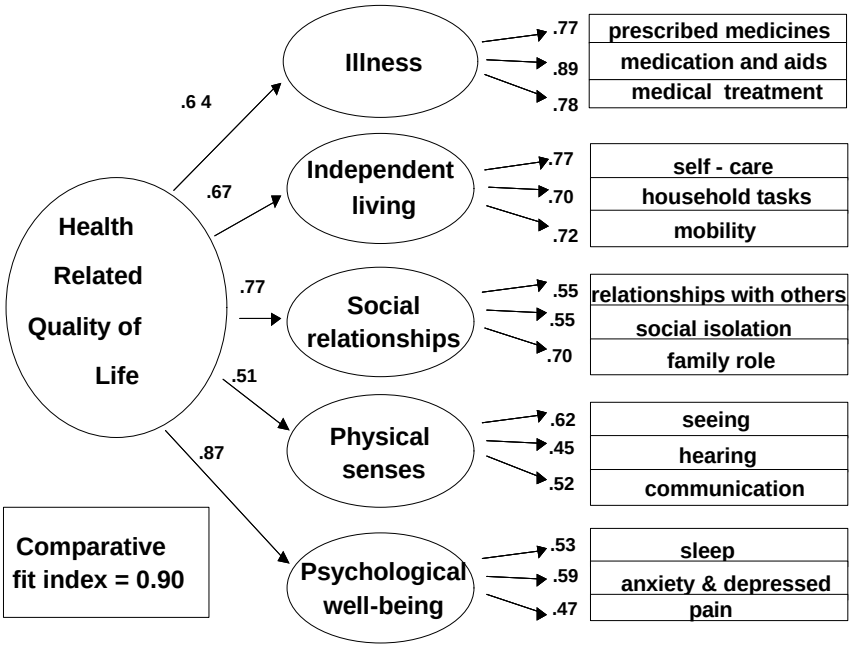
[6] From decision analytic theory there are three levels of preference independence and this should dictate the use of an additive, multiplicative or multi linear model (von Winterfeldt and Edwards 1986; Feeney et al 1996).

the clearly incorrect assumption that different individuals may have their quality of life very significantly reduced by a significant loss of utility arising from different dimensions. A further source of likely error is that some instruments have adopted utility scoring techniques – and specifically the rating scale – which probably do not measure utility (Richardson 1994).

The AQoL project was designed to overcome, as far as possible, each of these problems. Specifically the project sought to create an instrument where the descriptive system is :

- derived using correct psychometric procedures for instrument construction and hence achieves construct validity;
- sensitive to as much of the full universe of HRQoL as is practical;
- based upon structurally independent dimensions of health.

Figure 1 Structure of the AQoL



The achievement of these properties and the creation of the AQoL descriptive system is described elsewhere (Hawthorne et al 1997; 1999; Richardson and Hawthorne 2000). The procedures adopted in this part of the project resulted in an instrument which is unique in two respects: viz,

- it has a hierarchical descriptive structure in which structural independence is achieved between dimensions but not within dimensions. This permits greater sensitivity within dimensions. This is shown in Figure 1;
- a descriptive system which can claim to have construct validity, which increases confidence in the validity of the health state descriptions.

Additional project objectives were :

- to scale the instrument using a flexible utility model and an accepted technique for preference measurement;
- to achieve preference independence between dimensions;
- to achieve a valid trade-off between quality and length of life.

Preference independence was sought by the selection and content of items[7]. The achievement of this property was then assumed, as elsewhere (Feeney, Torrance et al 1996).

Scaling of the AQoL descriptive system – the calibration of item responses and their combination into a single numerical value – is outlined in Hawthorne et al (2000). The multiplicative model used in this exercise and its properties are outlined in Richardson and Hawthorne (2000). Two notable problems arose in this context. The first was the appropriate treatment of negative values derived from the time trade-off technique. In principle and in practice these are unconstrained and assume values as low as minus infinity. The second and related problem concerns the estimation of the utility score of the instrument 'all worst' health state[8]. The estimation procedure necessarily involves survey respondents placing a value upon a 15 dimensional health state which, for many, is worst than death. The cognitive task in combination with the existence of negative scores makes the treatment of this pivotal value problematical. Our treatment of the task is described in Richardson, Hawthorne (2000).

Confidence in an MAU or psychometric instrument depends, in part, upon the process of construction and calibration. In (larger) part it depends upon the demonstration of validity in a range of contexts. By June 2000 the AQoL had been adopted in 44 projects and information from a number of these is being analysed. Three interesting sets of results illustrating three facets of the validation process are illustrated in Table 2 and Figures 2 and 3. The first of these, arising from the Southern Health Care Network Coordinated Care Trial indicates that the AQoL has significant predictive power and a capacity to distinguish between patients requiring intensive and less intense medical care. The second study (Figure 2) illustrates the breadth of coverage of the health domain achieved by the AQoL as compared with the SF36, the most widely used psychometric instrument in the world. Figure 3 arises from a study of cochlear implantation illustrates the discriminatory power of AQoL following a particular intervention.

7 Preference independence indicates that the preference score for an item does not depend upon the level of another item, dimension or combination of items (see von Winterfeldt & Edwards 1993; Feeney, Torrance et al 1996).

8 The multiplicative model produces a score where 100 and 0 represent, respectively, the instrument all best and all worst values. These values must be converted to utility measured upon a scale where 100 and 0 represent normal or good health and death respectively.

Table 2 AQoL and Actual Patient Expenditures in the 18 month period after completion of the AQoL in the Southern Health Care Network Trial

AQoL Value	Mean Cost per Year (AUD$)	No of Cases	Relative Cost
-0.04 - 0.10	8,765	105	7.2
0.11 - 0.20	7,157	66	5.9
0.21 - 0.30	6,750	91	5.6
0.31 - 0.40	4,469	93	3.7
0.41 - 0.50	4,727	131	3.9
0.51 - 0.60	3,606	149	3.0
0.61 - 0.70	2,455	156	2.0
0.71 - 0.80	2,027	225	1.7
0.81 - 0.90	1,708	233	1.4
0.91 - 1.00	1,213	278	1.0

Source: Results calculated from Segal et al Evaluation of the Southern Health Care Network Coordinated Care Trial, Melbourne, CHPE, 2000.

Notes: Mean cost per year includes MBS, PBS, AHP, Nursing.

Figure 2 Concept Map of Health Domain for Back Related Illness
A. SF 36 items and back pain

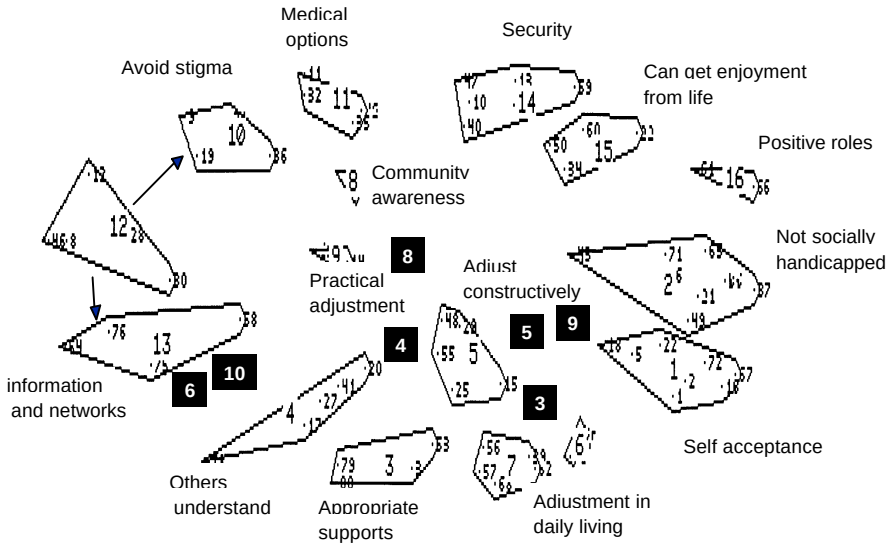


Figure 2 (contd) B. AQoL Items and Back Pain

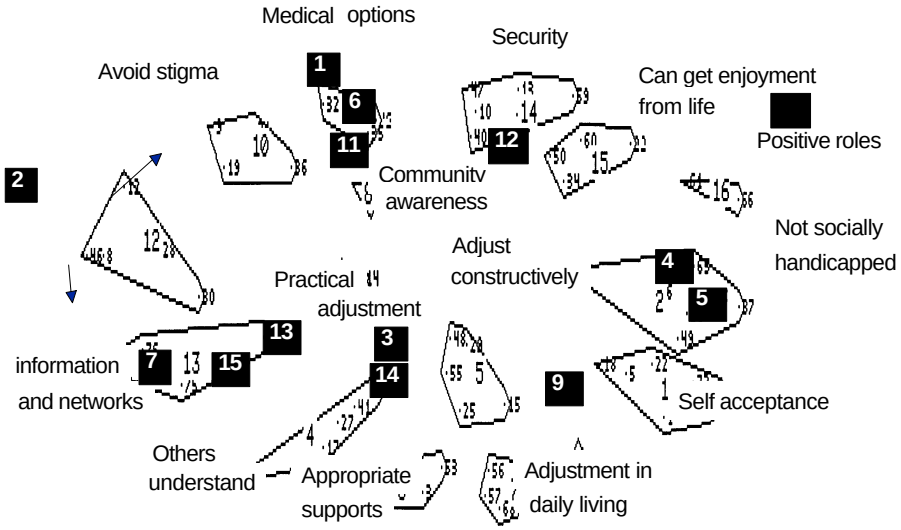
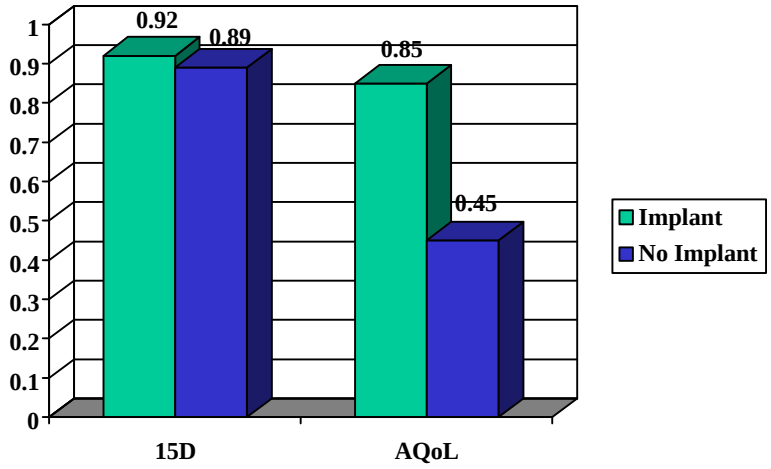


Figure 3 Result from the Cochlear Implant Study



3 The Validation Study

3.1 The survey

Six HRQoL instruments were administered to a stratified sample of Victorian residents, selected to cover a very broad range of health conditions from those who were healthy through to those who were terminally ill. The strata were: (a) randomly selected community members weighted by socio-economic status to achieve representativeness of the Australian population; (b) outpatients attending two of Melbourne's largest public hospitals (the method used was random sampling within selected timeframes); and (c) inpatients from three Melbourne hospitals (purposive sampling was used within wards based on severity of condition).

The six instruments were the SF-36 and WHOQOL-Bref (generic health status instruments) and the AQoL, EQ5D, HUI III and 15D (utility instruments). All instruments were scaled or scored as recommended by the developers. To avoid response bias instrument order was systematically rotated. This paper reports on the data analysis for the four utility instruments only.
A range of analyses were used, including scattergrams, correlations, analysis of variance and structural equation modelling.

3.2 The utility instruments

Each of the five utility instruments reported here consists of a 'descriptive system'; ie a series of item stems and responses which seek information about a concept or 'element' of the universe of HRQoL. Responses to these are then weighted and combined to produce the index.

For the AQoL, the descriptive system comprises 15 items, each with 5 reference categories. These are combined into 4 dimensions. The dimensions are Illness, Independent Living, Social Relationships, Physical Senses and Psychological Wellbeing (Hawthorne et al 1999). The utility weights were derived from an Australian population sample using time-trade off (TTO). During

the calculation of the utility index, the Illness dimension score is not used. A multiplicative function is used to combine the remaining four dimensions into the utility index (Hawthorne et al 2000).

The EQ5D (formerly the EuroQoL) consists of 5 items, each of which has 3 ordinal levels in the item responses. The items measure Mobility, Self-care, Usual Activities, Pain/Discomfort and Anxiety/Depression. The utility weights were obtained from a representative sample of the UK population, using the TTO. The utilities are computed using a regression model in which each item level is considered (Dolan et al 1996).

The HUI III comprises 15 items. The number of item responses varies between 4–6; again at an ordinal level. Of the 15 items, 12 are used in the utility score and form 8 'attributes'. These were constructed to be what the authors described as 'within the skin' attributes; that is, they focus upon disability and impairment rather than upon handicap. They are Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition and Pain. The utility weights were derived using a visual analog rating scale (VAS), the values of which were transformed based on valuations obtained from the standard gamble. The weights reflect those of the Canadian population. As with the AQoL, the HUI III uses a multiplicative model for combining the attributes into the index score (Furlong, Torrance and Feeney 1996; Furlong et al 1998).

The 15D consists of 15 items and, like the EQ5D, each item represents a dimension. The 15D also focuses primarily on 'within the skin' dimensions, covering Mobility, Vision, Hearing, Breathing, Sleeping, Eating, Speech, Elimination, Usual Activities, Mental Function, Discomfort & Symptoms, Depression, Distress, Vitality and Sexual Function. The weights used were from the adult Finnish population and were elicited using rating scales (Sintonen 1995).

The SF36 was designed as a psychometric and not, originally, as a utility instrument. Its 36 items place particular emphasis upon physical ability and vitality, general health and anxiety/depression. Utility weights have been created for the instrument using respondents to a standard gamble survey in the UK (Brazier 1998). Results here are based upon a preliminary algorithm provided by Brazier and may be amended at a future date.

3.3 Some issues

The five utility instruments reviewed here differ in virtually all respects. This makes direct comparability difficult. First, the 'perspective' on HRQoL differs. The EQ5D offers a very simple functional perspective. The HUI III and the 15D reflect a 'within the skin' perspective: that is items refer exclusively to impairment or disability: these instrument do not purport to measure handicap encountered in a social context, but impairment or disability to the contextless individual. The AQoL attempts to incorporate handicap and contains some questions probing the impact of impairment or disability upon a person's life and social functioning.

Second, the descriptive systems differ in the dimensions included and the number of items in each dimension. This is shown in Table 3 for the five utility instruments.

Third, the different instrument designers adopted different methods for weighting the instruments. The 15D was weighted using a rating scale; the EQ5D and AQoL the time trade-off (TTO) method, and the HUI III a rating scale which was then transformed into an estimate of standard gamble scores using a function fitted to selected health states for which both rating scale and

standard gamble scores were obtained. In addition, the time period for which health states were to be endured also differed. For the AQoL and EQ5D the health state duration was specified as 10 years, while for the HUI III the duration was a lifetime (defined as 60 years).

Table 3 HRQoL coverage: key instruments

HRQoL dimensions	SF -36	AQoL[2]	EuroQoL	HUI -III	15D
Relative to the body					
Anxiety/Depression	***	*	*		**
Bodily care	*	*	*	*	
Cognitive ability				*	*
General health	******				
Memory				*	
Mobility	***	*	*	*	*
Pain	**	*	*	**	*
Physical ability/Vitality	*******			*	*
Rest and fatigue	**	*			*
Sensory functions		**		****	*****
Social expression					
Activities of daily living		*	*		*
Communication		*		**	*
Emotional fulfilment	**			**	
Family role		*			
Intimacy/Isolation		*			
Medical aids use		*			
Medical treatment		**			
Sexual relationships					*
Social function	**	*			
Work function	**				

The coverage of the five instruments is summarised in Table 3.

Fourth, the method of computing the utilities also varied. The 15D uses an additive model in which final disutility scores are a weighted average of the disutility for each item. The rating scale weights for the relative importance of each dimensions are re-scaled so that the weights sum to unity. The AQoL and the HUI III use a multiplicative model in which a declining score on any dimension results in a fixed percentage decline in utility which remains after taking into account the disutility arising from the other dimensions. The EQ5D utilities are computed using a linear regression model derived from the econometric relationship between TTO scores for whole health states and the utility scores on each of the dimensions.

While, at first, it may appear that such diverse methods will inevitably result in very different estimates of health states utilities, this is not inevitable. It is possible to use quite dissimilar instruments to measure the same quantity. For example, physical weight may be measured using either a spring or balance scale; distance, temperature and other physical quantities are

commonly measured with different instruments employing different scales. Nevertheless, given the diversity of measurement strategies represented by the four utility instruments, disparate results would be unsurprising.

A source of error in the comparison of instruments may arise because of the definition of best possible health implied by the all-best health state in each instrument. In particular, as instruments become more sensitive or detailed at the upper end of the scale more sources of ill health are ruled out by the statement of the all-best health state and a utility score of 1.00 will have a different meaning on different scales. In comparisons which might be invalidated by this source of error we redefined a utility score of 1.00 as follows. For each person who self rated their health as 'excellent' or 'good' on the SF36 we calculated an average score for each instrument. Individuals with a predicted instrument score equal to this average or above were assigned a utility of 1.00. That is, in these comparisons the value of 1.00 was defined by the same group of respondents. The value of 1.00 in these comparisons therefore corresponds with 'good health' as self rated, and is not the score of the instrument all best. It is arguable that this – or even a more conservative definition of 'normal health – should be employed in cost utility analysis as the average patient is returned to normal and not to best possible health. Reported comparisons incorporating this adjustment are described as 'adjusted results' other results use unadjusted instrument utility scores.

4 Results

The response rates to the validation study were 58 percent (n=396) for the community sample, 43 percent (n=334) for outpatients and 68 percent (n=266) for inpatients. Details of participants are given in Table 4. This shows 50 percent of respondents were male, the mean age was 52 years, 75 percent were born in Australia, and 64 percent had attended either primary or high school. Forty-four percent were working in paid employment and 34 percent were retired. Sixty percent were married and 18 percent were single.

4.1 Average utilities and association between instruments

Table 5 reports the average utility score obtained from each instrument broken down by respondent status (in-patient, out-patient, community member) and by age categories. The community results are plotted against age in Figure 5. The broad pattern by respondent status is as expected. There is a decreasing gradient in scores between hospital and community respondents. There is the expected decreasing gradient in scores with a rise in age and between community respondents and inpatients in wards. Along both these dimensions the gradient is particularly pronounced for the AQoL and least pronounced for the SF36.

The distribution of scores is shown in Figure 4. This indicates that the frequency distributions and keratosis for the five instruments are quite different. AQoL and HUI result in a greater range of scores and assign lower values to more people than the other instruments. By contrast the SF36 and 15D have comparatively truncated distributions. Scores for the AQoL are generally lower than those for other instruments. This pattern is more obvious in Figure 5 which plots utilities from the five instruments for members of the community.

Table 4 Demographic and Other Characteristics of Respondents

Gender	Male	488	50%
	Female	488	50%
Age	Mean (sd)	52.4	(18.0)
Birthplace	Australia	731	75%
	Other	245	25%
Education level (a)	Primary	116	12%
	High	488	52%
	TAFE/Trade	127	13%
	University	216	23%
Employment status	Fulltime	300	31%
	Part time	126	13%
	Home duties	100	10%
	Student	30	3%
	Retired	328	34%
	Unemployed/Other	85	9%
Marital status	Single	175	18%
	Married/de facto	581	60%
	Separated/Divorced	105	11%
	Widowed	116	12%
Health Status	Hospital In-patients	142	16%
	Hospital Out-patients	333	38%
General Population	> 17 years	403	46%

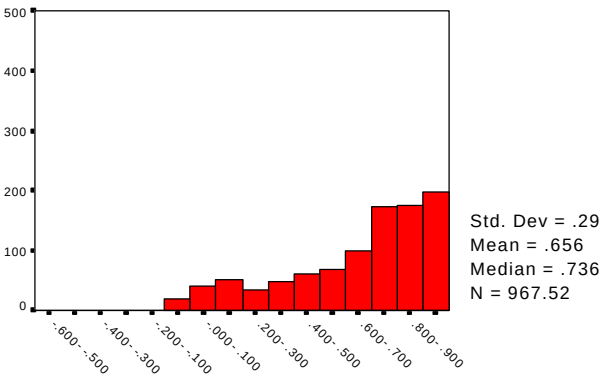
Notes: The number of missing cases for any variable can be computed by subtracting the table entries from the base of 996.

a = Highest level achieved

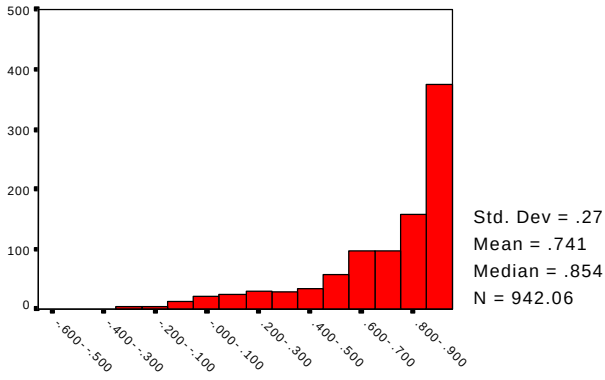
Table 5 Average Utilities by Age and Patient Status (Adjusted)

				AQOL3 adjusted	HUI-3 adjusted	15D adjusted	EUROQoL adjusted	SF-36 Utility adjusted
				Mean	Mean	Mean	Mean	Mean
Respondent type	Popln	Age in Years	16-35	.929	.938	.969	.944	.969
			35-50	.872	.903	.944	.901	.967
			50 - 65	.884	.904	.945	.894	.955
			65 - 95	.791	.827	.906	.823	.935
	Outpatient	Age in Years	16-35	.716	.806	.888	.739	.894
			35-50	.722	.793	.878	.747	.887
			50 - 65	.771	.781	.888	.787	.906
			65 - 95	.626	.638	.833	.704	.859
	Ward	Age in Years	16-35	.637	.772	.871	.649	.866
			35-50	.525	.692	.806	.488	.817
			50 - 65	.533	.668	.808	.563	.823
			65 - 95	.493	.595	.791	.580	.830

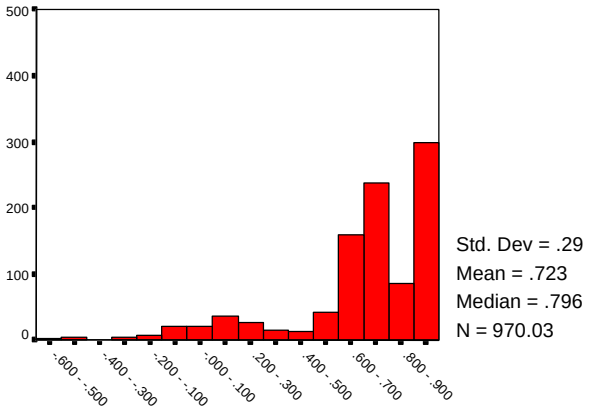
Figure 4 Distribution of Utility Scores



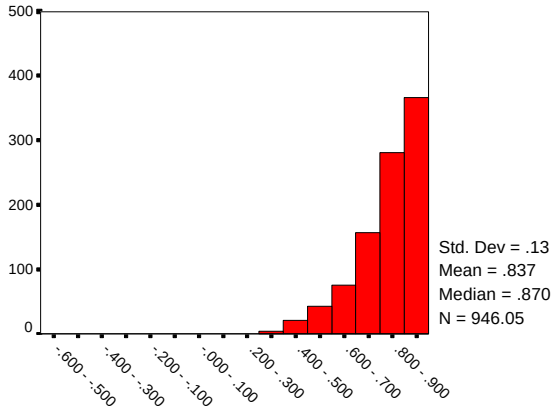
AQoL



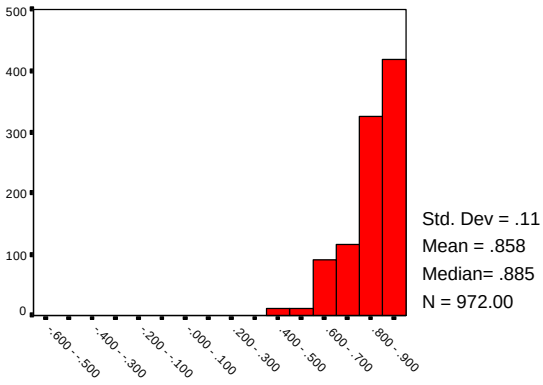
HUI-3



EUROQoL



15 D



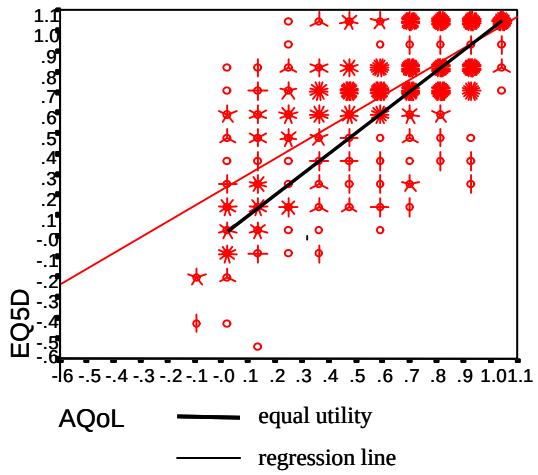
SF36 Utility

4.2 Variation and sensitivity

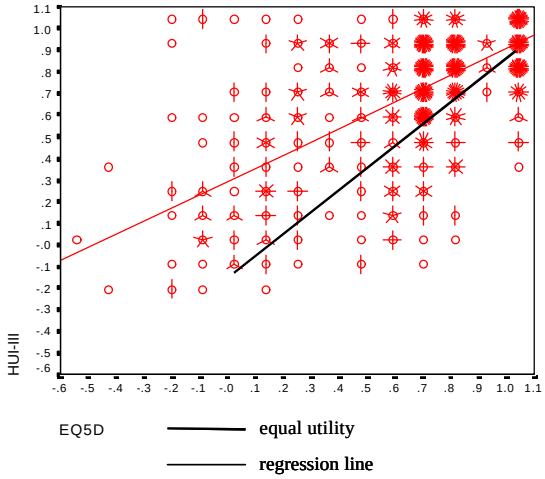
The relationship between individual utilities on the different instruments are plotted in Figure 6 which reveals extremely high levels of variation around the theoretical ideal of equal utilities. Such an ideal outcome would result in a scattergram which coincided with the 45^{o} line through the origin, ie in the equation: *instrument score (1) = instrument score (2)*. Two issues arise. The first is whether or not the variation is so great that at least one and possibly both of the instruments in each comparison is invalid and dominated by measurement error. The second is the extent to which results arise from the different sensitivities of the instruments in different domains of health.

Figure 5 Scattergrams of Instrument Scores

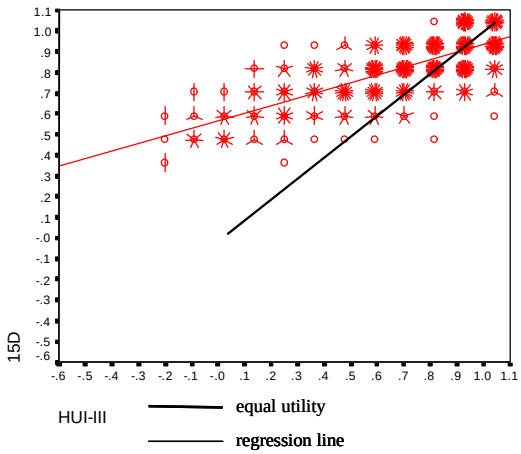
AQoL and EQ5D



EQ5D and HUI III



HUI III and 15D



AQoL and SF36

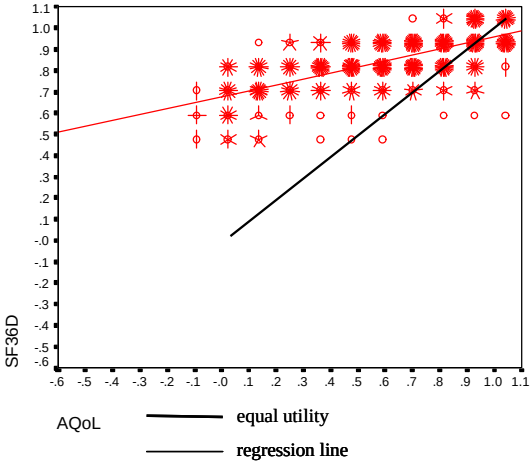


Figure 5 cont'd.

SF6D and 15D

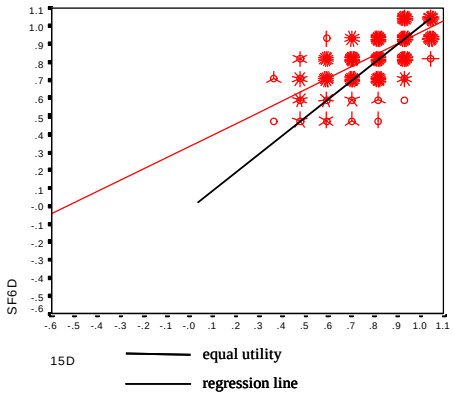
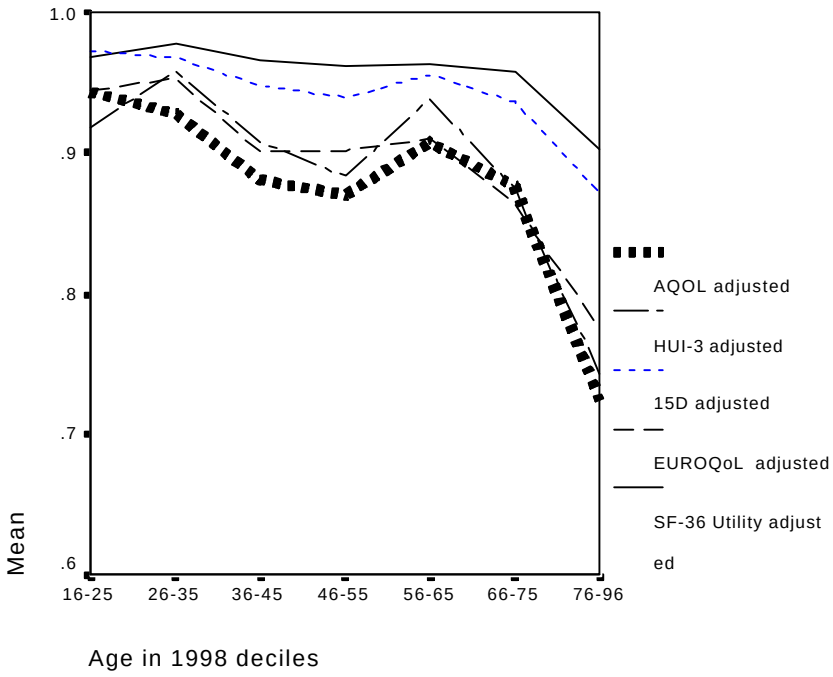


Figure 6 Mean QoL Score by Age; Community Sample, Adjusted Scores



Notes:
Community informants only, adjusted scores

It is normal for there to be a very large standard error in the measurement of health state utilities. This is not necessarily a lethal outcome in an analysis in which individual scores have little significance. Cost utility analysis requires average QALY values based upon average utility scores. The relevant test of consistency, therefore, is the average association between instrument scores. From the correlation coefficients reported in Table 6, it is clear that, despite the visual impact of Figure 6, there is a relatively high correlation between instrument scores. As judged by the standards of psychometric validation these coefficients lend support to the

hypothesis that the instruments are all measuring the same underlying concept. The average correlation of each of the instruments with other instruments is relatively high. The largest single correlation is between the AQoL and 15D and the highest average correlations are also obtained for these two instruments. The SF36 and HUI have the lowest average correlations although the absolute differences are small.

Table 6 Correlations between Instruments (Unadjusted)

	AQoL	HUI III	15D	EuroQoL	SF-36
HUI III	.762				
15D	.821	.799			
EuroQoL	.751	.653	.760		
SF36 utility	.733	.664	.741	.7.25	
Mean	.767	.715	.775	.722	.716

Notes:

Population, Outpatient and Ward cases, N = 906 - 968

A more powerful and important test of the association between instrument scores is the magnitude of the incremental change in the score of each instrument which results from an incremental change in the score of other instruments. In principle this is a more important test of the validity of an instrument than the absolute magnitude of the utility. This is because cost utility analysis deals with incremental improvements in HRQoL and, consequently, it is the magnitude of the incremental change that must be valid. Thus, a change in the true index of utility of 0.1 should be associated with a change in measured utility of 0.1. Alternatively, the slope of the linear relationship between true and estimated utilities should be unity.

The results in Table 7 which report these relationships are, for this reason, possibly the most important that we report in the paper. Table 7 reports the ratio of the incremental changes predicted by two instruments. This is equivalent to the 'b' coefficient in the linear relationship: *instrument (1) = a+ b instrument (2)*. This linear relationship may not, however, be obtained from regression analyses which assume that the independent variable is error free. An appropriate technique for overcoming this problem was suggested by Barnett (1969) who provided an algorithm for the calculation of the linear relationship when three variables were involved. The procedure was, in effect, a special case of what subsequently was generalised into simultaneous equation modelling procedures. We have employed these to generalise the Barnett method and to extend its use to the case of five variables.[9]

From Table 7, AQoL predicts a greater change in utility than any of the other instruments (coefficients in row one exceed unity). Predicted changes, however, are similar to those predicted by the HUI or EuroQoL. By contrast, 15D and, particularly, the SF36 predict very significantly lower changes in utility with the SF36 predicting only 35 and 42 percent of the

[9] Summarising, the five instruments are used to create a latent variable as discussed later and reproduced in Figure 10. This is then used as the independent variable in a relationship between each instrument score and the latent variable. Instrument scores may then be equated. Thus: if S1 … S5 are the five instrument scores, a latent variable L is calculated. A series of linear relationships are then calculated of the form $s_1 = a+bL$. From this

$$s_1 = a_1+b_1L \;;\; a_2+b_2L = s$$
$$s_1 = \cdot c+d\, S_2$$

where $c= a_1 – a_2\,(b_1/b_2)$

$d= (b_1/b_2)$

change obtained from AQoL and HUI (row five). The magnitude of these discrepancies indicates that while both sets of instruments (AQoL, HUI and EuroQoL vs 15D and SF36) may, (or may not) predict valid, reliable or even cardinal indices of HRQoL, one or the other or both does not have the 'strong interval property' required for the measurement of QALYS.[10]

Table 7 Average ratio of incremental changes (Generalised Barnett Coefficients)

	AQoL	HUI3	EuroQoL	15D	SF 36U
AQoL	1.000	1.194	1.105	2.182	2.856
HUI	.837	1.000	.925	1.827	2.391
EuroQoL	.905	1.081	1.000	1.975	2.585
15D	.458	.547	.506	1.000	1.309
SF36U	.350	.418	.387	.764	1.000

These average relationships do not reflect instrument sensitivity. That is, for any score on a given instrument there might be wide dispersion in utilities measured by a second instrument reflecting sensitivity to domains included in the second but not in the first instrument. Of course, it might, simultaneously, be true that at any given score in the second instrument the first would also vary because of greater sensitivity along another dimension. This possibility is investigated in Figure 7. This is constructed by identifying individuals with the best possible score on each of the instruments respectively and plotting the variation in these individuals utilities as measured by the other instruments. Instruments with greater sensitivity would be expected to demonstrate greater variation for these individuals.

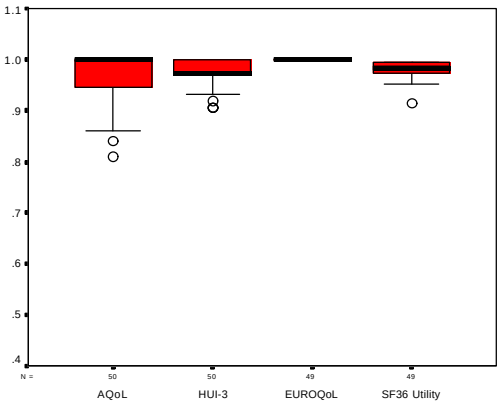
Results in Figure 7 suggest greatest variability in the AQoL and HUI. In contrast, the EQ5D and SF36 reveals virtually no variability when either the AQoL or 15D are at their all best level. Conversely, there is least variation in other instruments when AQoL, 15D and SF36 are at their maximum scores. Taking account of the compression of the 15D it reveals a surprisingly high variation when the EQ5D and HUI indicate full utility.

Of course results in Figure 7 also reflect random variation which, in principle, could account for all of the results. This hypothesis and the conflicting hypothesis that it is true sensitivity reflected in the results is tested, anecdotally, in two case studies reported in Figure 8. These suggest that it is the omission of domains of health and the relative importance of domains which accounts for the very large discrepancies in the reported scores in these two cases. Work is currently underway to investigate these issues in greater detail. A preliminary, more general test of sensitivity using self TTO is discussed later in this paper.

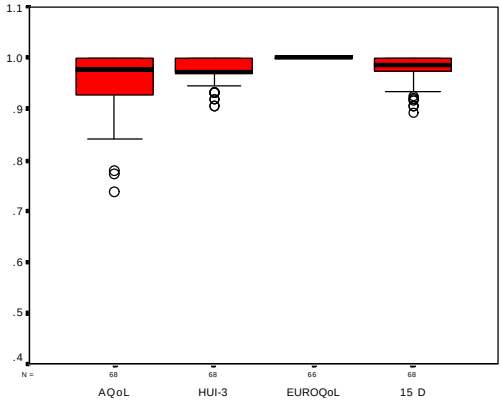
[10] This is the property that, a given percentage change in the numerical value of the index of HRQoL should be equally important (as judged by rater preferences) as an equal percentage change in the quantity of life (Richardson(1994)).

Figure 7 Distribution of Other Instrument Scores at Full Health (Adjusted)

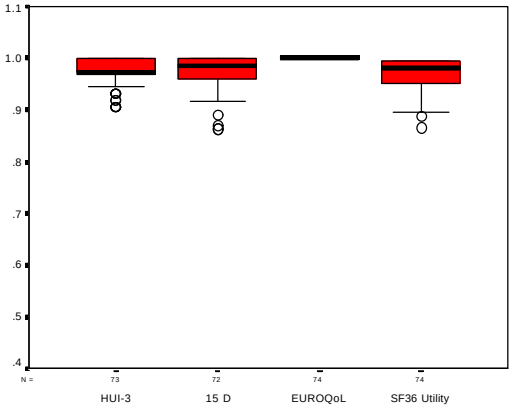
15D = 1.0



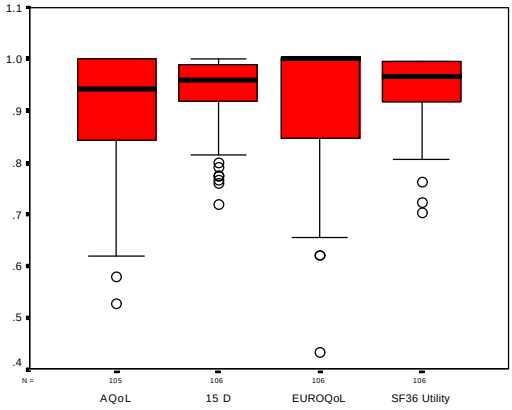
SF36 Utility >= 0.995



AQoL = 1.0: HUI



III = 1.0:



EuroQoL = 1.0:

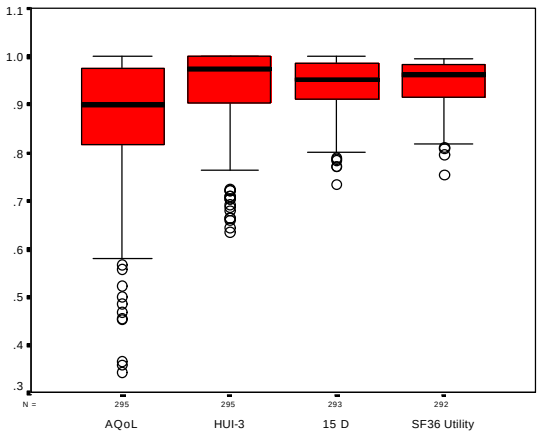


Figure 8 Two Case Studies

Case Study 1

Health Dimension	AQoL	15d
Physical health and mobility	• Gets around home/community without difficulty • Has some difficulty focussing. • Hears normally.	• Walks normally has slight difficulty • Cannot read text; can see to walk • Hears normally • Shortness of breath on exertion • Eats normally • Serious bowel/bladder problems
Activities of daily living	• Needs no help with personal care • Or with household tasks	• Performs usual activities without difficulty
Bodily pain, General Health	• Suffers severe pain • Sleeps in short bursts only: is awake most of the night	• Severe physical discomfort/pain • Has great problems with sleeping. • Feels very weary
Social function	• **Has no close warm relationships** • Has friends and is not lonely • **Some parts of the family role affected by health.** • No difficulty communicating	• Speaks normally • Sexual activity almost impossible ?
Emotional and mental health	• Moderately anxious worried or depressed	• Feels extremely sad and anxious • Slight difficulties with thinking and memory
SF-36: FAIR HEALTH	**0.14 (0.49)**	**0.55**

Case Study 2

Health Dimension	HUI-3	EuroQol
Physical health and mobility	• Walks without difficulty • Full use of hands and fingers • **Unable to see well even with glasses** • **Some hearing difficulty**	• No problems walking around ?
Activities of daily living	• Bathes, eats and dresses normally	• No problems with personal care • No problems performing usual activities
Bodily pain, General Health	• Moderate pain, occasionally disturbing normal activities • Health rated as fair	• Moderate pain or discomfort
Social function	• No problems with communicating	
Emotional and mental health	• Occassionally fretful, angry or depressed • Somewhat forgetful, but able to think clearly	• Not anxious or depressed
SF-36: AVERAGE HEALTH	**0.14 (0.74)**	**0.80**

4.3 Instrument structure

The validity of an MAU model depends, inter alia, upon the elimination of redundancy in an instrument; that is, the same element of a health state should not be measured more than once. In a structurally independent model correlation between items would be entirely attributable to their common correlation with the underlying concept and not as a result of their direct measurement of the same element[11]. Results of SEM confirmatory factor analyses of each instrument are reported in Figure 9. As judged by the comparative fit index which reflects the proportion of item variation attributable to the instruments structure, these results indicate that the EQ5D and the AQoL are the best performing instruments.

A similar analysis may be conducted using, not dimension or item results of an instrument as input, but, the utility scores from each of the five instruments. This analysis, reported in Figure 10, indicates the extent to which variation in a particular instrument's score is explained by the underlying concept of health related quality of life as defined by the five instruments taken together. This indicates that the AQoL is more closely associated with this global concept than the other instruments.

4.4 Self rated TTO: Preliminary Analysis

A small subset of 126 patients were asked to evaluate their current health state using the TTO instrument which was applied using the same 'flip flop' technique as employed in the construction of the AQoL and as described by Torrance (1986). A significant proportion refused to trade. Where these individuals indicated a low quality of life on other instruments we deleted them from the subsequent analysis. This was justified by the assumption that self rating would create a greater initial resistance to trade than the abstract rating of some other health state. One half of those who would not trade were retained in the analysis.

Self TTO scores were regressed on the utility scores predicted by each of the five generic instruments. Results shown in Table 8 indicate that the explanatory power of every instrument is low with the HUI III and EQ5D explaining less than 20 percent of variation. A comparison of the B coefficients in Table 8 again identifies the two separate groups of instruments. In the first group, the AQoL, HUI III and EQ5D have low B coefficients: a change in the predicted utility from these instruments corresponds with an increase in the TTO score of between 0.31 and 0.43. In contrast the 15D and SF6D have B coefficients above unity. And, in particular, the magnitude of the change in utility predicted by the 15D is almost identical to the magnitude indicated by the self TTO.

The self TTO data allow a further and more general test of instrument sensitivity. The residual from each of the five regression analyses reported in Table 8 were regressed against each of the remaining four instruments to determine whether any of these instruments had additional explanatory powers of the self TTO scores. This may be expressed as follows:

[11] Thus, for example, there might be correlation between depression and immobility: those who are depressed are unwell and this prevents their mobility. This is distinct from mobility and depression measuring the same element which they do not. Each of the two elements may independently contribute to a lower HRQoL. By contrast a correlation between immobility, incapacity to play sport and difficulty carrying out household tasks could all measure the sae underlying element and result in double counting.

$$TTO = \alpha_1 + \beta_i \text{ (Instrument 1)} + e$$
$$e = e_1 + e_2$$
e_1 = systematic
e_2 = random
$$e = \alpha_1 + \beta_i \text{ (Other instrument)} + e2$$

In sum, this analysis of residuals indicates whether or not the 'other instrument' which is the independent variable can explain variation in the TTO which is not explained by instrument 1.

Results are reported in Table 9. Shaded diagonal results are the B coefficients and R^2 scores for the regression of the TTO on each instrument, as reported in table 8. Other results refer to the regression of the error term of this equation against other instruments. Thus, for example, the residual from the regression of self TTO on the HUI III upon the AQoL (row 2, column 1) indicates that 6 percent of the error will be explained. Similarly, from row 3, 35 and 33 percent of the residual error from the regression of self TTO on the EQ5D can be explained by the 15D and SF6D respectively. In contrast, from rows 4 and 5 instruments in group 1 (AQoL, HUI, EQ5D cannot explain any of the residual error from the regressions for the 15D and SF6D. This suggests a greater sensitivity to variation in self TTO by these latter instruments than by the former three instruments. It must be emphasised that these are preliminary results and are based upon the little know self TTO and upon a very small group of patients.

Table 8 Regression Coefficients: TTO dependentn =

Predictor Variable	B Coeff	beta (Pearson Cor)	R Square
AQoL	0.429	0.452	0.204
HUI 3	0.386	0.407	0.166
15 D	1.096	0.538	0.289
EuroQol	0.308	0.345	0.119
SF6D Utility	1.344	0.488	0.238

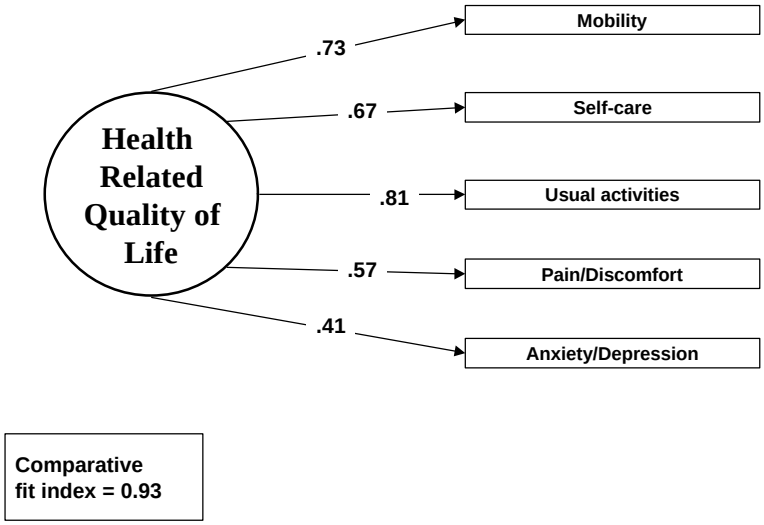
Table 9 Explanatory power of 4 instruments of the residual TTO from the regression of TTO on the 5th instrument

DEPENDENT: Residual from regression of TTO on:		**INDEPENDENT**				
		AQoL	**HUI**	**EQ5D**	**15D**	**SF6D**
AQoL	B	* 0.43				0.51
	R^2	0.20			0.22	0.23
HUI III	B	0.19	* 0.37	0.19	0.54	0.19
	R^2	0.06	0.17	0.06	0.10	0.06
EQ5D	B	0.20	0.29	* 0.31	0.62	0.77
	R^2	0.25	0.29	0.12	0.35	0.33
15D	B	ns	ns	ns	* 1.12	ns
	R^2				0.29	
SF36D	B	ns	ns	ns	ns	* 1.36
	R^2					0.24

* Regression TTO = a + b (Instrument)

Figure 9 Structural Equation Analysis of Individual Instruments

EQ5D



HUI III

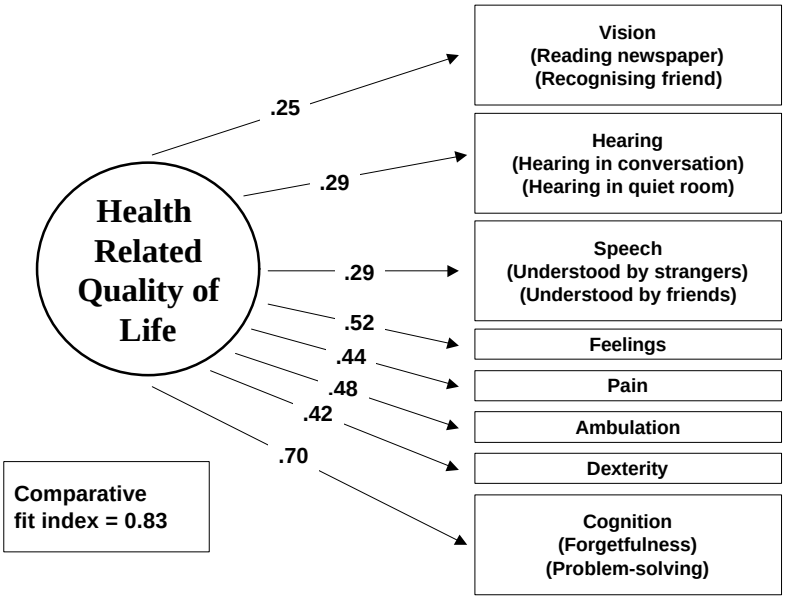
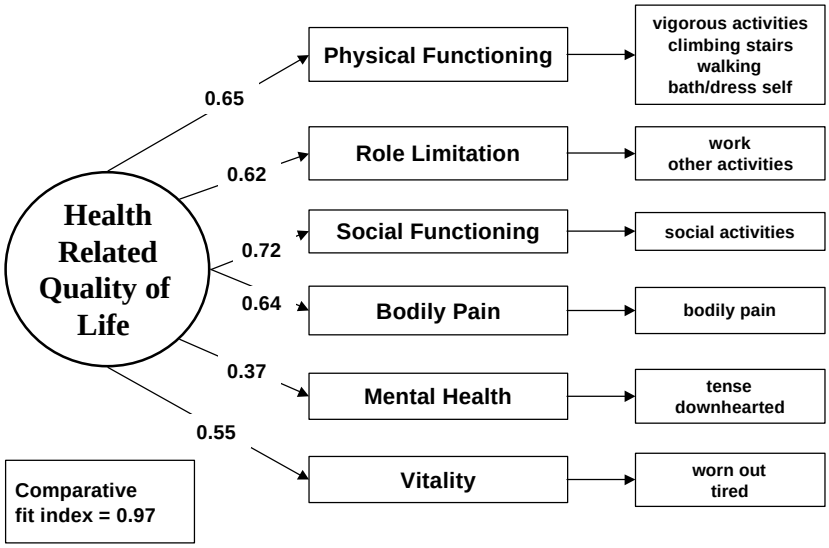


Figure 9 (contd.) **Structural Equation Analysis of Individual Instruments**

SEM of the SF6D



15D

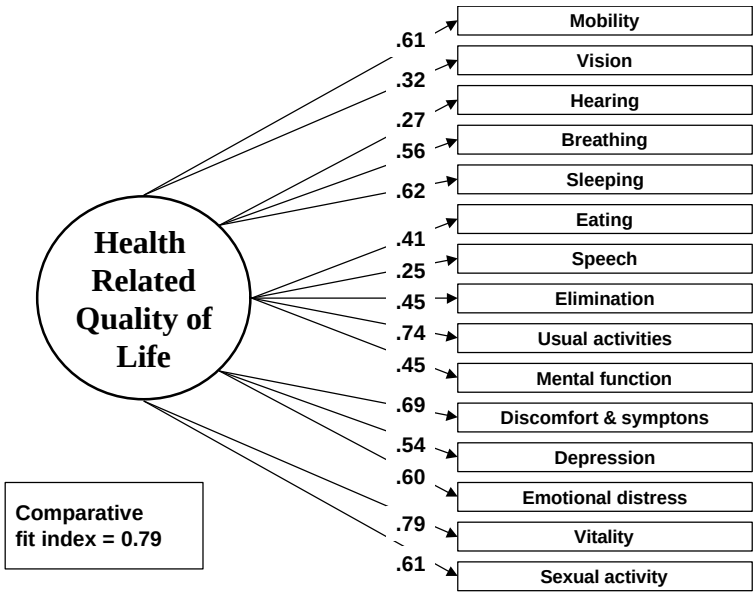


Figure 9 (contd.) **Structural Equation Analysis of Individual Instruments**

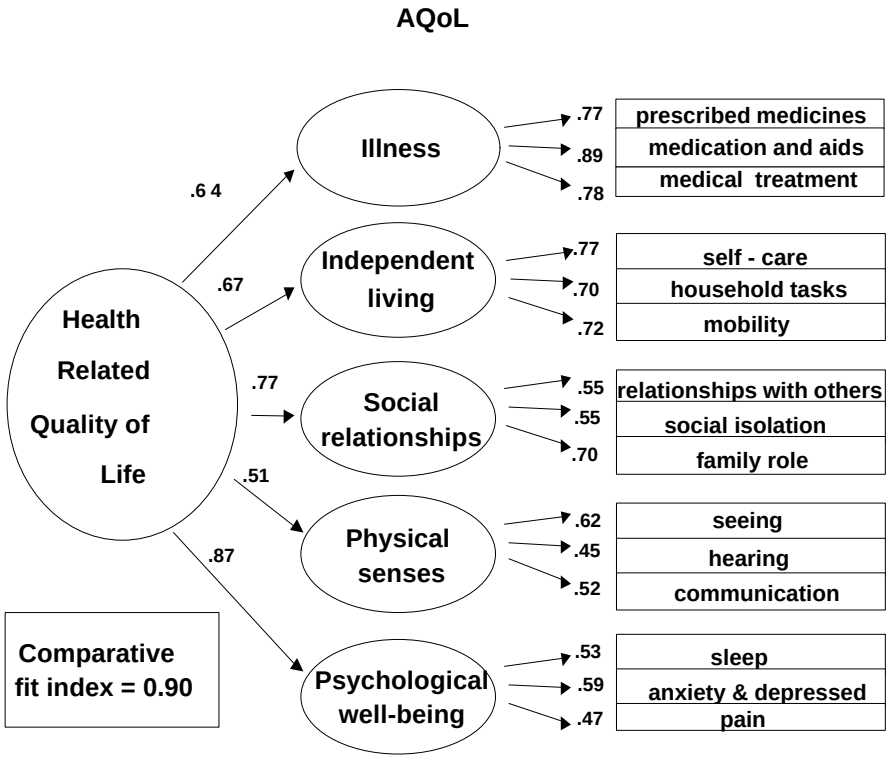
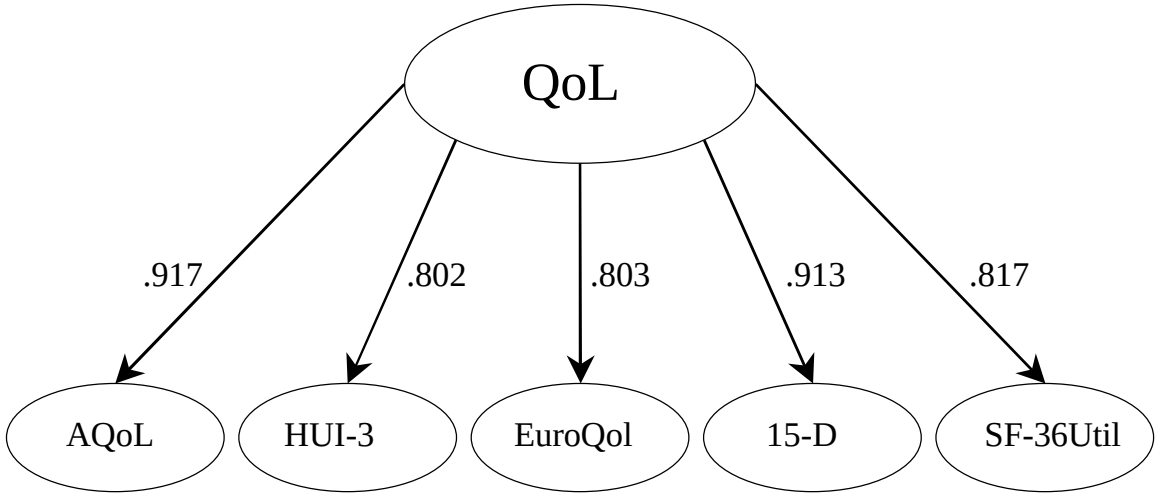


Figure 10 **Structure of Quality of Life Instruments (Congeneric Structural Model)**



Note:

CFI =.992, Fitted using EQS using adjusted utility scores, N=966

5 Discussion

Evaluation requires criteria and to this point the criteria for evaluating an instrument have not been made explicit. The paper has been concerned with the validation of the AQoL by its comparison with other widely used instruments. In doing so, of course, it also considers the validity of these instruments. The appropriate evaluative criteria, therefore, are those which establish instrument validity.

In general terms validity is defined as the measurement of what an instrument purports to measure. In the present context instruments purport to measure the magnitude of the 'utility' which is appropriate for the construction of quality adjusted life years (QALYs) where the defining property of a QALY is that it is equivalent to a year of full (or normal) health as judged by individuals. It is possible to separate two levels of this QALY property. First, the 'weak' QALY property may be defined by a metric having an interval property such that increments of the unit are of equal value. A necessary but not sufficient condition for this is that the units satisfy the various tests usually employed to establish psychometric validity; viz, appropriate correlation with other instruments or measurements which are known or believed to measure the desired property. Secondly, a 'strong QALY' property may be defined as a metric with the 'strong interval' property as defined by Richardson (1994). This implies that an x percent increase in the utility, as measured, is of equal value as an x percent increase in life years; that is the metric acts as the exchange rate between the quality and quantity of life which is the defining characteristic of the QALY.

This paper has been primarily concerned with the weak QALY property of the five instruments. This has been tested using a series of sub-criteria; viz, (i) the coverage of the relevant domain of HRQoL by the instrument's descriptive system; (ii) evidence that the instrument is measuring a commonly accepted concept of HRQoL; (iii) the sensitivity of an instrument to a change in the health state; (iv) the appropriate correlation between instrument scores; and (v) the quantitative relationship between incremental changes in instrument values.

With respect to the first two criteria the AQoL performs very well. It has a broad coverage of the different dimensions of health (face validity) and, as defined by changes in other instrument scores, it is sensitive to varying health states. Its internal structure, as investigated by confirmatory factor analysis, indicates that it measures differences in true health as defined by the totality of the instruments employed in the study. It is highly correlated with other instruments as required for psychometric validity, ie for the demonstration of the weaker QALY property. Results from the application of the Barnett procedure indicates that, on average, AQoL produces a change in utility scores between health states which is very similar to the score obtained from two of the most commonly used instruments, viz, the EQ5D and HUI.

Establishing the strong QALY property is more difficult. Sub-criteria include (e) preference independence; (f) non redundancy or structural independence; (g) correlation with people's directly stated preferences; (h) plausible results from the test of reflective equilibrium, ie that the implication of the utility scores for life death decisions elsewhere are plausible.

Violation of preference independence would imply that the importance of a particular item or dimension response depends upon the person's quality of life along some different item or dimension. Preference independence is usually assumed and is outside the scope of the present

study. However, gross violation of the property would result in perverse results from other comparisons. Structural independence is tested in large part by the confirmatory factor analyses. By this criteria the AQoL performs very well. This is unsurprising as the AQoL was constructed to achieve this property.

Results of the comparison with directly stated preferences – own TTO scores – is worrying. If own TTO scores were the gold standard the results would invalidate all of the instruments. However the status of own TTO is unknown and has seldom been investigated. It may be subject to a greater 'shock-horror' effect and require greater deliberation than the more abstract questions normally asked. Respondents may be sceptical about – or even unable to envisage – the option of returning to full health if they have not experienced this for many years. In sum, self TTO in its 'spontaneous' form (with inadequate time for deliberation) is a still largely unknown entity.

The present study has not investigated 'reflected' equilibrium. In a later study we will systematically investigate the relationship between instrument scores and the implication for life-death decision making.

6 Conclusion

As judged by the criteria above the AQoL instrument performs well and with respect to some criteria better than alternative instruments. It includes domains of health excluded by other instruments and displays considerable sensitivity. Utility scores perform as expected by a psychometric instrument and are consistent with the requirements of the weak QALY property.

Two unresolved issues are evident with the scores arising from the AQoL and, to a greater or lesser extent, from other instruments. First, individual utilities often vary very significantly between instruments and, even at the aggregate level there is very imperfect correlation. This should not be used as an argument for abandoning the use of generic MAU instruments and reliance upon unique, scenario based studies. While these may be warranted in a number of contexts it is important to recognise that individual scenarios are seldom validated and that weaknesses uncovered by explicit validation studies, such as the one reported here, may be matched, or worse, in the case of individual scenarios where comprehension, interpretation, cognitive overload, etc have an unknown affect upon results. Secondly, the absolute utility scores of AQoL and HUI are low and may violate the requirements of the strong QALY property. (If they do not then the other instruments studied fail this test). This property remains undemonstrated in all instruments and suggests that those using these instruments should not regard the implications for the trade-off between life and quality of life as having been conclusively resolved. That is, where appropriate both life years and quality adjusted life years gained as a result of an intervention should be reported.

The overall conclusion from this study is that the AQoL has been 'validated' with respect to the weak interval property. This does not imply that it is the appropriate instrument to use in every context. Rather, it has performed well as judged against other instruments when subjected to a limited range of tests. The evidence presented here is strong enough to justify the use of the AQoL. However two other major conclusions of the study are, first, that either the AQoL, HUI III and EQ5D or the SF6D and 15D do not have the strong interval property. Secondly, the descriptive systems of the instruments differ very widely in their coverage of different domains of health related quality of life and the differences in utility scores found at the individual level is

certainly attributable, in part, to these differences. For this reason our strongest recommendation to those seeking a generic MAU instrument is to select the one which is most sensitive to the health states in which they are interested. Because of the variability noted above it is highly desirable that users should include more than one generic instrument in their study. At the present state of instrument development this is the most effective way of carrying out an effective sensitivity analysis.

References

Brazier J et al 1998, 'Deriving a preference-based single index from the UK SF-36 health survey', *Journal of Clinical Epidemiology,* 51:1115–1128.

Dolan P, Gudex C et al 1996, 'Valuing health states: A comparison on methods', *Journal of Health Economics*, 15:209-231.

Torrance G, Furlong W, Feeny D, Boyle M 1995, 'Multi-attribute preference functions: health utilities index', *PharmacoEconomics*, 7:503-520.

Furlong W, Feeny D, Torrance G, Goldsmith C, DePauw S, Zhu Z, et al. 1998, *Multiplicative Multi-attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report*. Working Paper. Hamilton: McMaster University, Centre for Health Economics and Policy Analysis; 98-11.

Feeney D, Torrance G et al 1996, 'Health utilities index', *Quality of Life and Pharmacoeconomics in clinical trials,* B Spilker, Lippincott-Raven Publishers, Philadelphia.

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, Melbourne.

Hawthorne et al 1999, 'The assessment of Quality of Life (AQoL) instrument: a psychometric measure of Health Related Quality of Life (HRQoL)', *Quality of Life Research*, 8:209-224.

Hawthorne G, Richardson J et al 2000, 'Construction and Utility Scaling of the Assessment of Quality of Life (AQoL) Instrument', Working Paper 101, Centre for Health Program Evaluation, Melbourne.

Kaplan R, Ganiats T et al 1996, 'The quality of well-being scale', *Medical Outcomes Trust Bulletin,* 4:2-3.

Kind P 1996, 'The EuroQoL instrument: an index of health-related quality of life', *Quality of Life and Pharmacoeconomics in clinical trials,* B Spilker, Lippincott-Raven Publishers, Philadelphia.

Murray C & Lopez A 1996, *The Global Burden of Disease*, World Health Organization, Geneva.

Richardson J 1994, 'Cost utility analysis: what should be measured?', *Social Science and Medicine,* 39:(1)7-21.

Richardson J & Hawthorne G 2000, 'Negative Utility Scores and Evaluating the AQoL All Worst Health States', Working Paper No 73, Centre for Health Program Evaluation, Melbourne.

Rosser R 1993, 'A health index and output measure', *Quality of Life Assessment: Key Issues in the 1990s,* S Walker and R Rosser, Kluwer Academic Publishers, Dordrecht.

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

Sintonen H 1995, *The 15D measure of health-related quality of life: feasibility, reliability and validity of its valuation system*. Melbourne: National Centre for Health Program Evaluation; Working Paper 42.

Torrance G 1986, 'Measurement of health state utilities for economic appraisal: a review', *Journal of Health Economics,* 5:1-30.

Williams A 1995, 'The measurement and valuation of health: final report on the modelling of valuation tariffs', MVH Group, Centre for Health Economics, University of York, York.

von Winterfeldt DV & Edwards W 1986, *Decision Analysis and Behavioural Research*, Cambridge University Press, Cambridge.