

19517787**Request # 19517787****MAR 17, 2006****Mail To:**

VA MEDICAL CENTER
 CLINICAL INFORMATION MANAGEMENT-LIBRARY
 Attn: Mary Hess
 216 S. FOSTER DR.
 BATON ROUGE, LA 70806

DOCLINE: Journal Copy Epayment

Title: Annals of medicine.
 Title Abbrev: Ann Med
 Citation: 2001 Jul;33(5):358-70
 Article: A comparison of the Assessment of Quality of Life
 Author: Hawthorne G; Richardson J; Day NA
 NLM Unique ID: 8906388 Verify: PubMed
 PubMed UI: 11491195
 ISSN: 0785-3890 (Print) 1365-2060 (Electronic)
 Publisher: Taylor & Francis, Stockholm :
 Copyright: Copyright Compliance Guidelines
 Authorization: MEH
 Need By: N/A
 Maximum Cost: **Free**
 Patron Name: Dr. Dumitrescu (mihnea@med.va.gov)
 Referral Reason: Not owned (title)
 Library Groups: FreeShare,VALNET,EFTS
 Phone: 1.225.761-6850
 Fax: 1.225.761-6805
 Email: mary.hess@med.va.gov
 Routing Reason: Routed to TXUDAF in Serial Routing - cell 2
 Received: Mar 17, 2006 (03:07 PM EST)
 Lender: DARNALL US ARMY COMMUNITY HOSPITAL/ FT HOOD/ TX
 USA (TXUDAF)

This material may be protected by copyright law (TITLE 17,U.S. CODE)

Bill to: LAUVNO

VA MEDICAL CENTER
 CLINICAL INFORMATION MANAGEMENT-LIBRARY
 Attn: Mary Hess
 216 S. FOSTER DR.
 BATON ROUGE, LA 70806

MFE 17 Mar 06
MEH

A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments

Graeme Hawthorne¹, Jeff Richardson² and Neil Atherton Day²

As part of the validation of the Assessment of Quality of Life (AQoL) instrument comparisons were made between five multiattribute utility (MAU) instruments, each purporting to measure health-related quality of life (HRQoL). These were the AQoL, the Canadian Health Utilities Index (HUI) 3, the Finnish 15D, the EQ-5D (formerly the EuroQoL) and the SF6D (derived from the SF-36). The paper compares absolute utility scores, instrument sensitivity, and incremental differences in measured utility between different instruments predicted by different individuals. The AQoL predicted utilities are similar to those from the HUI3 and EQ-5D. By contrast the 15D and SF6D predict systematically higher utilities, and the differences between individuals are significantly smaller. There is some evidence that the AQoL has greater sensitivity to health states than other instruments. It is concluded that at present no single MAU instrument can claim to be the 'gold standard', and that researchers should select an instrument sensitive to the health states they are investigating. Caution should be exercised in treating any of the instrument scores as representing a trade-off between length of life and HRQoL.

Keywords: cost-utility analysis (cost-effectiveness analysis); economics; quality of life; utility.

Ann Med 2001; 33: 358-370.

Introduction

In benefit-cost analysis the ('opportunity') cost of an activity is defined as the value of the commodities, services or programmes that cannot be provided if resources are used to carry out that activity. Con-

sequently, economic evaluation attempts to compare benefits gained and lost from the use of resources in a particular way. In the health sector this comparison has proved to be particularly difficult. Costs are usually measured in dollars (reflecting the market value of benefits foregone), and health benefits include the value of human life and the quality of that life. Neither of these benefits is easy to measure in dollars. However, despite this, some form of comparison must be undertaken if resources are to be used wisely.

Cost-effectiveness analysis (CEA) attempts to circumvent this difficulty by measuring benefits in 'natural' units, such as the number of lives or life years saved. The criterion for selecting a programme or project is that it minimizes the 'cost' or 'foregone benefit' of obtaining a life or life year. In the last 30 years, economists have extended this approach to include health-related quality of life (HRQoL) in the calculation of 'benefits'. In cost-utility analysis costs are compared with quality-adjusted life years (QALYs), which quantify the strength of people's preferences for health states as defined by HRQoL measures. This concept of the strength or intensity of preferences is what economists mean by 'utility'.

The quantification of utility requires two broad tasks. First, the health state under investigation must be described. Second, a scaling technique, such as time trade-off (TTO) or standard gamble (SG), must be used to obtain a numerical value for the health state. This value should measure the strength of a person's preference for the health state.

Three approaches to this two-stage procedure have been used; holistic (or 'composite') measurement, the direct valuation of personal health, or the use of a multiattribute utility (MAU) instrument (1). With the holistic approach a scenario or vignette is constructed which describes the health state. The vignette is then 'scaled' by using a survey to elicit 'utility' values for the health state. With the direct valuation approach, respondents can be asked to evaluate and scale their current health state. In practice this approach has seldom been used. With the MAU approach a generic

From the ¹Centre for Health Program Evaluation, Department of Public Health, The University of Melbourne, and the ²Centre for Health Program Evaluation, Monash University, West Heidelberg, Vic, Australia.

Correspondence: Graeme Hawthorne, PhD, Centre for Health Program Evaluation, PO Box 477, West Heidelberg, Victoria 3081, Australia. E-mail: graeme@unimelb.edu.au, Fax: +61 3 94964424.

'descriptive system' or 'descriptive instrument' is created which consists of a series of health-related items and responses. Different health states may then be described by using different combinations of responses. Utility weights are then attached to every possible response. Although in principle all the health states can be scaled (as with the original Rosser-Kind Index (2)) this is impractical when there are a large number of states. Consequently, each of the major instruments has used a limited number of health state values to calibrate a model from which other values have been inferred. The model may be derived from the econometric analysis of the observed utilities, or may use decision analytic techniques to fit a simple additive or multiplicative model. Once scaled, a MAU instrument may be used to estimate the utility of all possible health states described by the descriptive system of the model.

To date, only a handful of generic instruments have attempted to measure utility; ie, the UK Rosser-Kind Index (2), the US Quality of Well-Being (QWB) (3), the Canadian Health Utility Index (HUI) instruments (4-7), the Finnish 15D (8-10) and the European EQ-5D (formerly the EuroQoL) (11, 12). Brazier has provided a utility scoring algorithm for the SF6D, derived from the SF-36 (13); and, finally, working for the World Bank and the World Health Organization, Murray and Lopez have published 'disutility' weights for the different health states required for the construction of disability-adjusted life years (DALYs) and

Key messages

- There are substantial differences between the five utility instruments with respect to their conceptual models of health-related quality of life, the content of the descriptive systems, the methods of weighting the different levels of health status, the algorithms for combining the different items and dimensions into utility scores and the range of theoretical utility scores available. Despite these differences, the instruments produce utility scores that are remarkably consistent.
- Caution should be exercised in treating any of the instrument scores as representing a trade-off between life-length and health-related quality of life.
- Because of their different properties, researchers should select the utility instrument that is sensitive to the health states which they are investigating.

used these to quantify the burden of disease in every country in the world (14). The Assessment of Quality of Life (AQoL) instrument developed by the present authors is the most recent MAU instrument (15, 16).

Table 1. Properties of the major utility instruments.

Scale	Coverage*	Type of description†	No of dimensions	Valuing method‡	Psychometric properties		Combination model	Instrument boundaries
					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
HUI1	X	Impairment	4	TTO	No	No	Multiplicative	-0.21-1.00
HUI2	✓	Impairment/ disability	7	VAS/SG	No	Yes	Multiplicative	-0.03-1.00
HUI3	✓✓	Impairment	8	VAS/SG	No	Yes	Multiplicative	-0.36-1.00
EQ-5D	X	Impairment/ disability	5	TTO	No	No	Regression/ Additive	-0.59-1.00
DALY	XX	Disease	N/A	PTO	No	No	RS/PTO [¶]	N/A
WHOOoL-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 (17). Coding scheme: XX = very poor, X = poor, ✓ = good, ✓✓ = very good.

† Based on WHO classification of diseases and impairments (18).

‡ ME, magnitude estimation; VAS Visual analogue scale; TTO, time trade-off; SG: standard gamble; PTO, person trade-off

§ Descriptive system constructed following standard psychometric rules for instrument construction (19, 20).

^{||} 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'.

¶ Rating scale validated by using the PTO. See text for full names of the instruments. N/A, not applicable; RS, rating scale.

The key characteristics of these instruments are summarized in Table 1.

Generally, MAU instruments have received little critical attention. There are, however, numerous unresolved issues. Some of these, such as the choice of a scaling method (SG, TTO, etc) are common to both of the broad approaches (holistic and MAU). Others are generic measurement issues, such as the need for construct validity. Some issues, however, are unique to the MAU approach, including the sensitivity of the 'descriptive system' to changes in health states; the achievement of preference and structural independence¹ that the model used to combine preference scores must represent the structure of people's true preferences and issues relating to the inclusion and interpretation of negative utility scores (21). Each of these issues has the potential to 'invalidate' a MAU instrument, ie, its numerical scores will not reflect 'true' utility. While there is no gold standard for evaluating instruments, it is possible to compare the scores that are predicted for the same health states. Large discrepancies in predicted values indicate that one or both of the instruments compared must be invalid: where two instruments are claiming to measure the same construct, they cannot both give valid predictions if these predictions differ. The present paper investigates the two issues of greatest immediate interest to the uses of MAU instruments, ie, the numerical values of utilities predicted by different instruments and the relationship between these values and the sensitivity of different instruments. As discussed above, the first of these issues relates to instrument validity. The second is concerned with the relevance of an instrument in the context of a particular health intervention.

Our analyses of these two broad issues are based upon the results of a large-scale survey designed specifically to validate the AQoL instrument through its comparison with four other instruments. To our knowledge, this is the largest comparative study of utility instruments to date.

The 'Assessment of Quality of Life' (AQoL) instrument

While each of the instruments reviewed above and presented in Table 1 has particular strengths, to our knowledge none was constructed by using normal psychometric principles to ensure construct validity

¹Preference dependence occurs when the strength of preference for some attributes in the descriptive system depends upon a person's health state as described by other attributes in the descriptive system. For a discussion of the three types of preference dependence, see Feeny et al (11). Structural dependence means that some element of a person's health state is described more than once. That is, there is 'double counting'.

and structural independence. Consider, for example, this second issue. MAU theory postulates that there should be no 'redundancy' among items in a descriptive system. That is, a single attribute should not be measured in more than one way (22). If redundancy occurs then, the (dis)utility of the attribute will be double-counted. However, the requirement of non-redundancy appears to conflict with the need for 'sensitivity', and several instruments have reduced redundancy by the adoption of very simple descriptive systems. However, this simplicity may have been achieved at the expense of sensitivity. A sufficient (but not necessary) condition for nonredundancy is that different scales within an instrument are orthogonal². This is usually demonstrated by establishing statistical independence through factor analytic techniques.

The achievement of these properties and the creation of the AQoL descriptive system is described elsewhere (23). In summary, the procedures adopted during the construction of the AQoL resulted in an instrument which is unique in two respects: 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 Fig 1); a descriptive system which can claim to have construct validity, which increases confidence in the validity of the health state descriptions.

Additional AQoL project objectives were: 1) to scale the instrument by using a flexible utility model and an accepted technique for preference measurement; 2) to achieve preference independence between dimensions; 3) to achieve a valid trade-off between quality and length of life. Preference independence was sought by the selection and content of items. The structural independence was obtained through the use of factor analysis during the creation of the descriptive system.

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 (16). Two notable problems arose in this context. The first was the appropriate treatment of negative values derived from the TTO. In principle and in practice TTO scores are unconstrained and can assume values as low as minus infinity. The second

²Orthogonality is not strictly necessary as scales may be 'environmentally correlated'. Von Winterfeldt and Edwards illustrate this in the case of a manufacturing plant, the management of which is concerned with the cost 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 (22). 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.

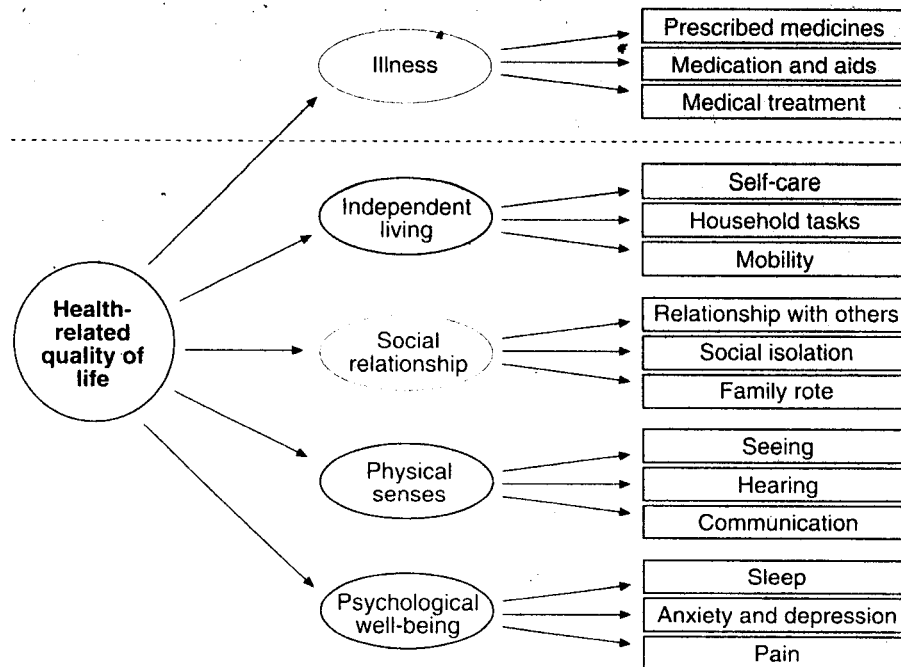


Figure 1. Structure of the Assessment of Quality of Life (AQoL) instrument. Note: The illness dimension is not used during the computation of utilities. (Adapted from (15, 16) with permission.)

and related problem concerns the estimation of the utility score of the instrument 'all worst' health state'. The estimation procedure necessarily involved survey respondents placing a value upon a 15-dimensional health state, which for many was worse than death. The cognitive task in combination with the existence of negative scores made the treatment of this pivotal value problematical. Our treatment of the task is described in the work of Richardson and Hawthorne (21).

Confidence in a MAU or psychometric instrument depends, in part, upon the process of construction and calibration. In 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, calculated from a randomized trial of the effect of funds pooling and service co-ordination, indicates that the AQoL has very significant predictive power and a capacity to distinguish between patients requiring intensive and less intense medical care. The second study (Fig 2)

illustrates the breadth of coverage of the health domain achieved by the AQoL when compared with the SF-36, the most widely used health status instrument in the world. Figure 3 shows data from a study of cochlear implantation and illustrates the discriminatory power of the AQoL.

Table 2. Assessment of Quality of Life (AQoL) and actual patient expenditures in the 18-month period after completion of the AQoL in the Southern Health Care Network Coordinated Care 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

Notes: Mean cost per year includes costs from the Australian Medical Benefits Schedule (covering doctor costs), Pharmaceutical Benefits Scheme (covering all prescribed pharmaceuticals), all other allied health professional costs (eg, physiotherapist) and all nursing costs. (Results calculated with permission from (24); unpublished data).

The multiplicative model produces a score where 1.00 and 0.00 represent, respectively, the instrument's all best and all worst values. These values must be converted to utility measured upon a scale where 1.00 and 0.00 represent good health and death, respectively.

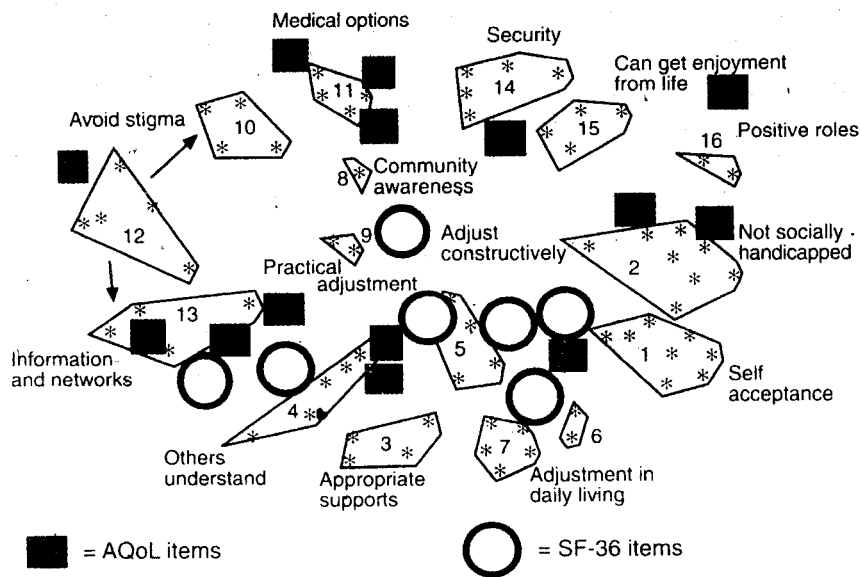


Figure 2. Concept map of health dimensions for back injury-related illness: comparing coverage of the SF-36 and AQL. Each asterisk = a statement; each number = a cluster of statements. Adapted from (16) with permission.

Methods

Six HRQoL instruments were administered to a stratified sample of residents in Victoria, Australia, 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: 1) randomly selected community members weighted by socioeconomic status to achieve representativeness of the Australian population; 2) outpatients attending two of Melbourne's largest public hospitals (the method used was random sampling within selected timeframes); and 3) inpatients from three Melbourne

hospitals (purposive sampling was used within wards based on severity of condition).

The six instruments were the SF-36 (26) and WHOQOL-Bréf (27) (both generic health status instruments) and the AQL (15, 16), EQ-5D (11, 12), HUI3 (5, 7) and 15D (8-10) (all utility instruments). The SF6D was derived from the SF-36 responses (13). All instruments were scaled or scored as recommended by instrument developers. To avoid response bias, instrument order was systematically rotated. This paper reports on the data analysis for the five utility instruments only.

The utility instruments

Each of the five utility instruments consists of a 'descriptive system' and a corresponding set of 'utility' values or a formula for deriving these values. As reported in Table 1, which shows a summary of the properties of each instrument, the instruments differ in their choice of descriptive system, the valuation method used to scale the instrument, the combination model and the range of utility scores predicted.

As noted, the AQL is the most recently developed. It is unique in having the hierarchical structure shown in Figure 1. The descriptive system consists of 15 items, each with four response categories. These are combined into five dimensions: illness, independent living, social relationships, physical senses and psychological well-being. The utility weights were derived from an Australian population sample by using the TFO. During the calculation of the utility index, the illness dimension score is not used. A

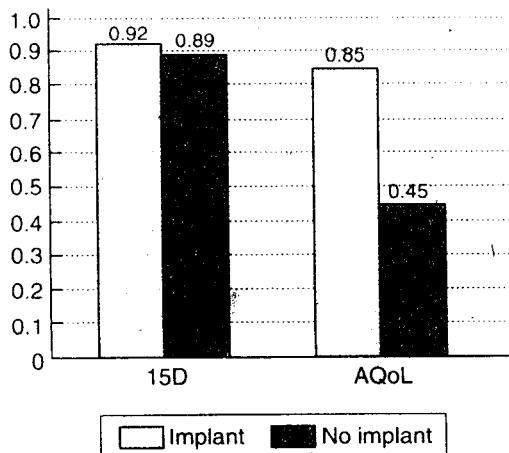


Figure 3. Result from the Cochlear Implant Study. (Calculated with permission from unpublished data from (25).)

multiplicative function is used to combine the remaining four dimensions into the utility index (16).

The EQ-5D (formerly the EuroQoL) consists of five items, each of which has three response categories. The items measure mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Utility scores were derived for each of the health states described by the descriptive system using econometric methods. Holistic utility scores for a selection of the health states were obtained by using the TTO technique. The TTO values were elicited from a British sample. These were regressed upon item response categories and the best fitting model used to predict the utility scores of the remaining health states (28).

The HUI3 consists of 15 items each with 4–6 response categories. Twelve of these contribute to eight attributes, which then combine to form the utility score. These attributes have a 'within the skin' focus; that is, they are primarily concerned with impairment rather than disability or handicap. The attributes are vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain. The utility weights were derived by using a visual analogue rating scale (VAS), the values of which were transformed to approximate SG scores by using a transformation function. This reflected the best fitting values obtained for four key health states. The weights reflect those of the Canadian population. As with the AqoL, the HUI3 uses a multiplicative model for combining the attributes into the index score (4, 7).

The 15D consists of 15 items, and like the EQ-5D, each item represents a dimension. The 15D focuses upon impairment and disability, and includes mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort and symptoms, depression, distress, vitality and sexual function. The weights were obtained from the adult Finnish population and were elicited with a VAS and then combined by using a simple additive model (8, 10).

Finally, the SF6D is based on the SF-36, which was designed as a psychometric instrument. It uses 12 of the 36 items to form six subscales; ie, physical function, role limitation, social function, bodily pain, mental health and vitality. The utility weights were derived by using the SG in a UK survey (13). The results presented here are based upon Brazier's preliminary algorithm, which may be amended at a future date.

As summarized in Tables 1 and 3, the five utility instruments differ in virtually all respects. First, the conceptualization of HRQoL differs, as shown in Table 1. The EQ-5D offers a simple functional description. The HUI3 and the 15D reflect a 'within the skin' perspective; that is, items mainly refer to impairment or disability; they do not purport to measure handicap encountered in a social context.

Table 3. Health-related quality of life (HRQoL) coverage by five utility instruments*

HRQoL dimensions†	SF6D	AQoL	EQ-5D	HUI3	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

Note: *Table shows only those items used in calculation of utility scores. Each asterisk represents an item. Based on item content examination.

†Dimensions of HRQoL defined by a review of 14 HRQoL instruments, 1971–1993 (17).

The AQoL focuses upon handicap but also contains some questions relating to impairment and disability. Second, the descriptive systems differ in the dimensions included and the number of items in each dimension. Only two dimensions are covered by all five instruments: mobility and pain (Table 3). Third, different scaling techniques are used (see Table 1). Fourth, the time period in the imaginary health state described to interview respondents differed. For the AQoL and EQ-5D the health state duration was specified as 10 years, while for the HUI3 the duration was a lifetime (defined as 60 years). Fifth, the model used to combine item scores varies between instruments (Table 1).

While, at first, it may appear that such diverse methods will inevitably result in very different estimates of health state 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 either with a spring or balance scale; distance, temperature and other physical quantities are commonly measured with different instruments employing different scales. Nevertheless, utility is a latent psychometric concept and, given the diversity of measurement strategies represented by the five utility instruments, disparate results would be unsurprising.

Results

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

Average utilities and association between instruments

The distribution of scores is shown in Figure 4, which reveals that the frequency distributions for the five instruments are quite different. The lowest negative utilities recorded were -0.04 (AQoL), -0.59 (EQ-5D); and -0.26 (HUI3). As shown, these three instruments provide a greater range of scores and assign lower values to many more people than the 15D and SF6D.

Table 5 reports the mean utility score obtained from each instrument broken down by respondent status (inpatient, outpatient, community member) and by age categories. The pattern by respondent status was generally in the direction expected: a monotonic decreasing relationship between sample (community, outpatient, inpatient) and age group within each sample (16–35, 36–50, 51–65, 66+). Scores should not be lower for the community sample than for the outpatient sample, and scores should not be lower for the outpatients than for inpatients. Within each sample, those aged 16–35 should not obtain lower scores than those aged 36–50, etc. For the AQoL and

Table 4. Demographic and other characteristics of respondents.

	<i>n</i>	Percentage (%)
Gender		
Male	488	50
Female	488	50
Age		
Mean (sd)	52.4	(18.0)
Birthplace		
Australia	731	75
Other	245	25
Education level*		
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 inpatients	142	16
Hospital outpatients	333	38
General population†	403	46

Notes: The number of missing cases for any variable can be computed by subtracting the table entries from the base of 996.
*Highest level attended. TAFE, Technical and further education.
†>17 years.

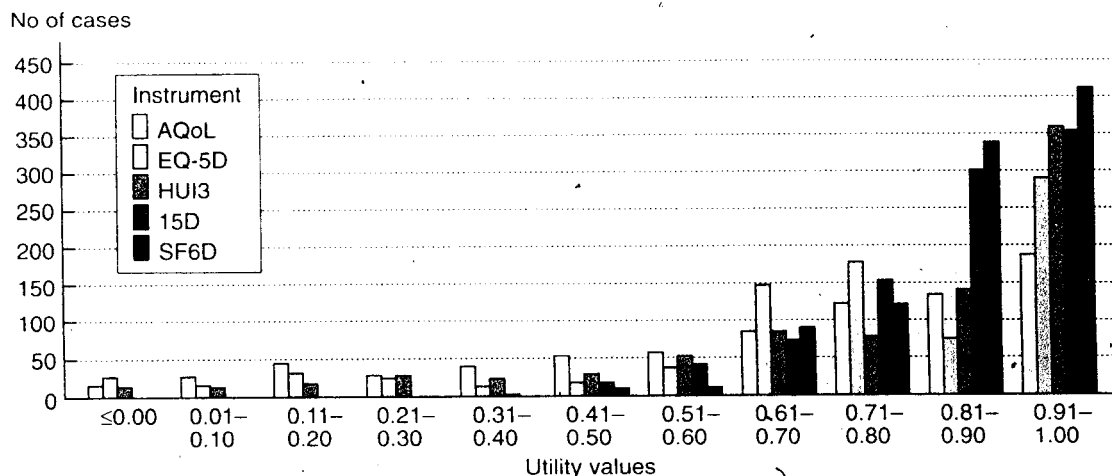


Figure 4. Distribution of utility scores.

Table 5. Mean utilities by age and patient status.

Cohort	Age group	Utility instrument				
		AQoL	EQ-5D	HUI3	15D	SF6D
Community sample	16-35	0.84	0.92	0.89	0.92	0.93
	36-50	0.80	0.88	0.86	0.91	0.93
	51-65	0.77	0.84	0.84	0.88	0.91
	66+	0.71	0.79	0.77	0.85	0.89
Outpatient sample	16-35	0.64	0.69	0.75	0.83	0.85
	36-50	0.64	0.71	0.74	0.82	0.85
	51-65	0.67	0.70	0.72	0.83	0.86
	66+	0.57	0.69	0.61	0.79	0.82
Inpatient sample	16-35	0.56	0.62	0.73	0.82	0.82
	36-50	0.47	0.47	0.65	0.76	0.78
	51-65	0.47	0.52	0.61	0.76	0.78
	66+	0.43	0.52	0.55	0.74	0.78

Note: The numbers in each age cohort were: 16-35: $n = 202$, 36-50: $n = 262$, 51-65: $n = 221$, 66+: $n = 285$.

SF6D there was just one case where mean scores did not follow this monotonic relationship, for the HUI3 and 15D there were two cases, and for the EQ-5D there were four cases.

Table 6 reports the correlation between each pair of instruments. This reveals that the two instruments most highly correlated were the AQoL and 15D, and that the two instruments with the least correlation were the EQ-5D and HUI3. The 15D and AQoL had, respectively, the highest and second highest average correlation with the other instruments. Overall, however, the instruments were well correlated with each other.

Despite this, the selected scatterplots in Figures 5a-e (which show only half the possible scatterplots) reveal some important differences that the correlations conceal, particularly when the data points are compared with the theoretically ideal relationship that would occur if both instruments predicted the same utility value; ie, the straight line marked on each scatterplot, which passes between the points (0.00, 0.00) and (1.00, 1.00). This ideal relationship ($\beta = 1.00$) can be used to assess the extent to which instruments differ in change scores.

This comparison revealed that the five instruments divide into two quite distinct groups. First, the

scatterplots between the AQoL, EQ-5D and HUI3 approximately correspond with the theoretical relationship. Similarly, the scatterplot of the SF6D and 15D also corresponds with the ideal relationship. However, the scatterplots between these latter two instruments and the AQoL, EQ-5D and HUI3 revealed a significant and systematic deviation from the ideal. Utility scores predicted by the SF6D and 15D are compressed into the upper range of the utility scale (a result determined by the instrument boundaries reported in Table 1). The important consequence of this is that a change in the average score reported by the SF6D or 15D corresponds with a much larger change in the scores predicted by the other three instruments. Geometrically this is shown in Figure 5c and 5d by the shallower slope of the line of best fit between the instruments.

Instrument sensitivity

There is no general test for instrument sensitivity, that is, for the capacity of an instrument to detect a change in health status. Sensitivity is different from the range of scores (as shown in Figures 5a-e above) and discriminant validity (ie, an instrument's ability to discriminate between cases in different health states; see Table 5). An instrument may be capable of detecting very small changes in health status but, if the numerical change in the instrument's score is small, then the range of values predicted may also be numerically small. Further, sensitivity may be context specific. Instrument A may be more sensitive to health states associated with one disease than instrument B, whereas instrument B may be more sensitive to the health states associated with a second disease.

In spite of these caveats the survey data permitted a relatively powerful test of sensitivity for each instru-

Table 6. Spearman correlations between utility scores.

	AQoL	EQ-5D	HUI3	15D	SF6D
EQ-5D	0.73				
HUI3	0.74	0.64			
15D	0.80	0.76	0.74		
SF6D	0.74	0.75	0.66	0.77	
Mean	0.75	0.72	0.70	0.77	0.73

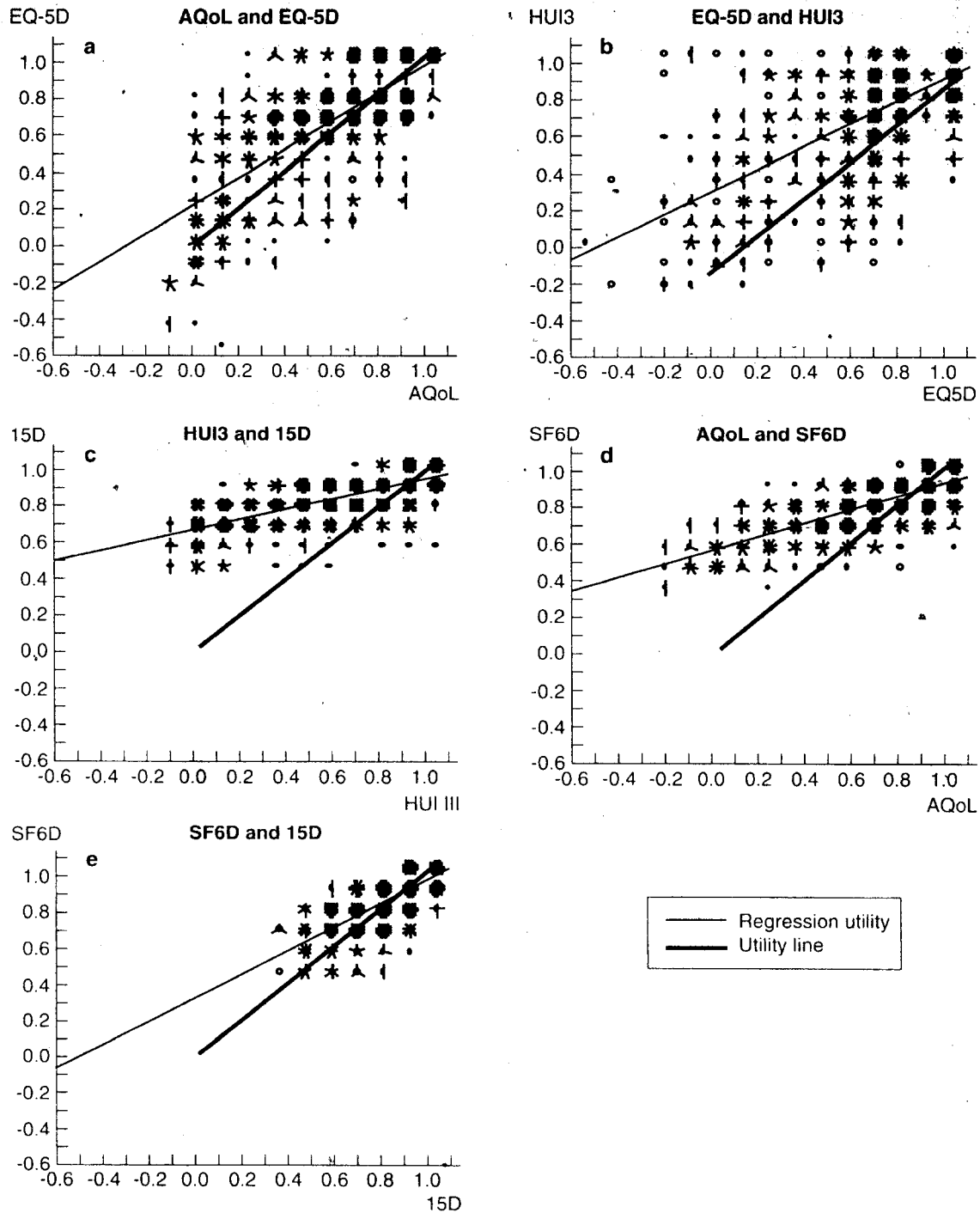


Figure 5a-e. Selected scatterplots of instrument scores. Note. In the interests of comparison, all instrument scores have been plotted on the same x- and y-axes. For the actual range of scores provided by each instrument, see Table 1.

ment. It is possible to determine whether or not other instruments detect changes in the health status when a given instrument does not. The results from this test procedure are reported in Figure 6 a-e. This was constructed by identifying those individuals with the 'all-best health state' (value = 1.00) on each of the

instruments in turn and plotting the variation in obtained utilities on the other instruments. In this test, the 'all-best health state' of an insensitive instrument would correspond with a wide dispersion of scores on the comparator instruments. Following T-score transformation to control for the different instrument

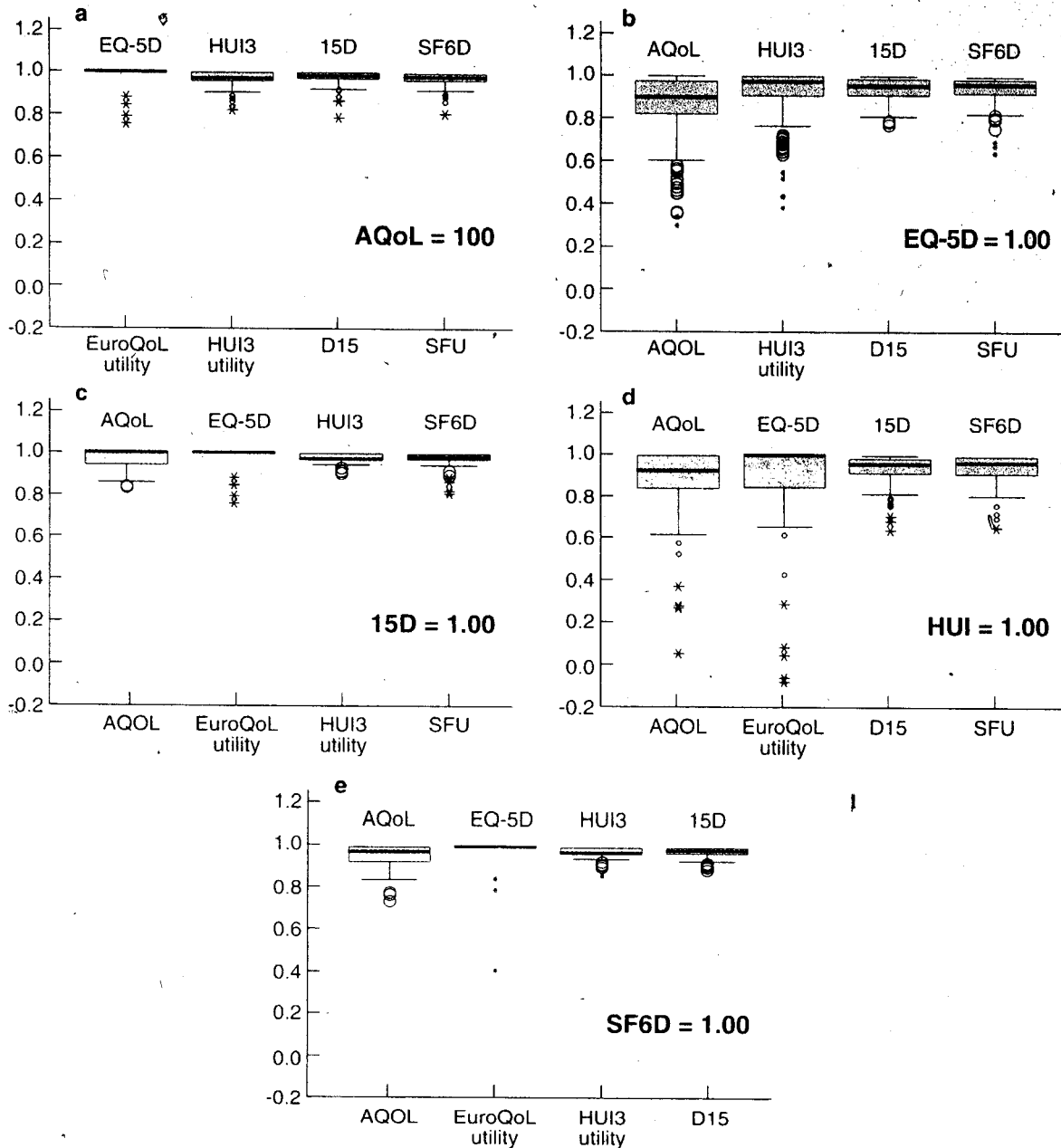


Figure 6. Instrument sensitivity as measured by instrument selection. The figure shows the effect of instrument selection, where for each instrument we identified those cases scoring the 'all-best health state' (value = 1.00) and then plotted the scores obtained on the other instruments. See the text for a discussion.

ranges, the mean departure scores (ie, variation) across instruments were calculated when the instrument of interest was held at '1.00'. This gave the following results where the lower the transformed scores the greater the departure from 1.00 for the other instruments. When the AQoL was held constant at '1.00', the mean departure score for the other instruments was 59.4, for the SF6D 59.2, the 15D 58.6, the HUI3 56.9 and the EQ-5D 56.6 (ANOVA, $F = 6.57$, $P < 0.01$). Figure 6 indicates that on this test the EQ-

5D and HUI3 were the least sensitive instruments; when a utility value of '1.00' was obtained on these two instruments the other instruments identified a wide dispersion in health status scores for these respondents. The greatest variability (ie, sensitivity near full health) was observed in the utilities of the AQoL, 15D and SF6D. Despite the compression in the ranges of the 15D and SF6D, Figure 6 reveals that these instruments detect differences in health status when the EQ-5D and HUI3 indicate full health.

Of course the results in Figure 6 also reflect random variation, which, in principle, could account for all of the differences. This hypothesis and the alternative hypothesis that the figure represents true sensitivity is tested, anecdotally, in the two case studies reported in Table 7. The value in the parentheses at the bottom of each of the HUI3 and AQL indicate what the utility would be if the 'bold' item responses were set at 'full health'. These suggest that it is the omission of dimensions of health and the relative importance of these dimensions that account for the very large discrepancies in the reported scores. Work is currently ongoing to investigate these issues in greater detail.

As a second test of sensitivity, scores from four instruments were used to predict whether or not an individual would have a score of unity on the remaining instrument. The predictive power is reported in the last column of Table 8. To do this, cases

were recoded into dichotomous values, such that those reporting 'full health' (1.00) were assigned and those reporting less than full health (≤ 0.99) were not assigned. Each instrument was then compared, iteratively, with the other four instruments. The results show the average predictive power of each instrument for correctly identifying those cases reporting a utility value of '1.00'. This table neatly classifies the instruments into two groups: those correctly predicting more than 50% of cases (the AQL, 15D and SF6D) and those predicting less than 50% of cases (the EQ-5D and HUI3). The results for the EQ-5D are particularly interesting as they suggest that the instrument has assigned far too many cases to a value of '1.00'; this may imply that the EQ-5D is insensitive at the top end of the range and cannot discriminate adequately between those with full health and those with some health problems.

Table 7. Case studies.
Case study 1

Health dimension	HUI3	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	Occasionally 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

Case study 2

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 Shortness of breath on exertion Eats normally Serious bowel/bladder problems
Activities of daily living	Needs no help with personal care	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
SF-36: Fair health	0.14 (0.49)	0.55

Table 8. Predicting against other instruments*

	Full health (n) [†]	Test sensitivity	Test specificity	False positive rate	Predictive power [‡]
AQoL	70	0.35	0.96	0.49	0.51
EQ-5D	290	0.86	0.74	0.79	0.21
HUI3	101	0.37	0.92	0.63	0.37
15D	58	0.31	0.97	0.42	0.58
SF6D	63	0.32	0.97	0.47	0.53

Notes: *Prediction of obtaining a value of '1.00' when compared with each of the other instruments iteratively. Mean values obtained and reported.

[†] Defined as > 0.99, assumed to represent 'full health'

[‡] Based on the standard formula for a 2 x 2 table: $pp = A/(A+B)$

Discussion

The paper has been concerned with the validation of the AQoL by its comparison with other widely used instruments. In doing so it also considers the validity of these instruments. In general terms validity is defined as the extent to which an instrument measures what it purports to measure. MAU instruments purport to measure the magnitude of the 'utility', which is appropriate for the construction of QALYs where the defining property of a QALY is that it is equivalent to a year of full health as judged by individuals. This implies two forms of interval property in the utility scale (29). First, a 'weak interval' property implies that increments of the unit are of equal value. For example, a movement from a utility of 0.20 to 0.40 must have the same value as the movement from 0.70 to 0.90. A necessary, but not sufficient, condition for this is that the units satisfy the various tests usually employed to establish psychometric validity, including appropriate correlation with other instruments or measurements that are known or believed to measure the desired property. Second, a 'strong interval' property implies that an 'x' percent increase in the utility, as measured, is of equal value as an 'x' percent increase in life years. Both of these changes have the same quantitative effect upon the number of QALYs and must, therefore, be of equal value. Consequently the strong interval property is necessary if the QALY is to combine correctly the quantity and quality of life.

This paper has been primarily concerned with the weak interval property of the five instruments. This has been tested by using a series of subcriteria: 1) the coverage of the relevant dimensions of HRQoL by the descriptive system of the instrument; 2) evidence that the instrument is measuring a commonly accepted concept of HRQoL; 3) the sensitivity of an instrument to a change in the health state; and 4) the appropriate correlation between instrument scores. 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), as defined by changes in other instrument scores, it is sensitive to varying health states, and it is highly correlated with other instruments.

Establishing the strong QALY property is more difficult. Subcriteria include: 5) preference independence; 6) nonredundancy or structural independence; 7) correlation with people's directly stated preferences; and 8) plausible results from the test of reflective equilibrium, ie, that the implication of the utility scores for life-death decisions elsewhere are plausible. These issues have received little or no attention in the literature and are the subject of current research. The present paper does, however, contain evidence relevant to the test of reflective equilibrium, and this represents the greatest threat to the validity of the AQoL. As shown in both the frequency distribution (Fig 4) and the scatterplots (Figs 5a-e), there are two quite distinct groupings of instruments; first, the AQoL, HUI3 and EQ-5D, and second, the 15D and SF6D. In contrast with the full life-death utility score ranges in group one, scores in group two are compressed into the upper half of the utility scale. Pair-wise comparison indicates that the change in the utility scores within each of the two groups of instruments is of comparable magnitude. However, changes in the utility scores in group one instruments correspond with changes of only half this numerical value in group two instruments. Both groups cannot be correct simultaneously. Either the AQoL, HUI3 and EQ-5D predict excessively low utility scores, or the 15D and SF6D predict excessively high utilities; that is, one group violates the strong interval property.

Conclusion

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.

A major conclusion is that the descriptive systems in the instruments studied here differ very widely in their coverage of different dimensions of HRQoL and that the reported differences in utility scores are attributable, in part, to these differences. Because of the variability noted above, our strongest recommendation is that it is highly desirable that users should include more than one generic instrument in any study. In addition, we would recommend to those seeking a generic MAU instrument to use in a particular context to select those that are most sensitive to the health states in which they are interested. Because of the major caveat concerning

the strong interval property, users of generic instruments, and especially those in the first group, should subject their utility results to rigorous sensitivity analysis.

We would like to acknowledge the generous cooperation and support of St Vincent's Hospital and the Austin and Repatriation Medical Centre, particularly the ward and outpatient staff who were so helpful to our interviewers. Our thanks are especially extended to Ms Helen McNeil, who coordinated data collection. We also extend our thanks to our interviewers: Sonia Barthelmebs, Kerith Culley, Janet Day, Marina Hawthorne, Ivan Marsetti, Jan Memery, Noelle Perry and Dr Press. Their high level of

professionalism and their willingness to work at times to suit the respondents all contributed to the success of this study. We would also like to thank respondents for completing consent forms and then interviews and questionnaires that were very long and challenging because the questions probed those parts of their lives that were, for many respondents, associated with the losses accompanying illness. This research was funded by the Victorian Health Promotion Foundation, to whom our thanks are extended. Without their generous support this research would not have been possible. Ethics approval was given by the Ethics Committees at Monash University, St Vincent's Hospital and the Austin and Repatriation Medical Centre. Dr Hawthorne's position at the University of Melbourne is funded by the Victorian Consortium for Public Health, to whom our thanks are also extended.

References

1. Torrance G. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986; 5: 1-30.
2. Rosser R. A health index and output measure. In: Walker S, Rosser R, eds. *Quality of life assessment: key issues in the 1990s*. Dordrecht: Kluwer Academic Publishers; 1993.
3. Kaplan R, Ganiats T, Sieber W, Anderson J. The Quality of Well-being Scale. *Medical Outcomes Trust Bulletin* 1996; 2: 3.
4. Torrance G, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions: health utilities index. *Pharmacoeconomics* 1995; 7: 503-20.
5. Feeny D, Torrance G, Furlong W. Health utilities index. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*. 2nd edn. Philadelphia, PA: Lippincott-Raven Publishers; 1996.
6. Feeny D, Furlong W, Torrance G. *Health Utilities Index Mark 2 and Mark 3 (HU12/3) 15-item questionnaire for self-administered, self-assessed usual health status*. Hamilton, Ont: Centre for Health Economics and Policy Analysis, McMaster University; 1996.
7. Furlong W, Feeny D, Torrance G, Goldsmith C, DePauw S, Zhu Z, et al. *Multiplicative Multi-attribute utility function for the Health Utilities Index Mark 3 (HU13) system: a technical report*. Working Paper. Hamilton, Ont: McMaster University, Centre for Health Economics and Policy Analysis; 1998: 98-11.
8. Sintonen H. *The 15D measure of health-related quality of life: feasibility, reliability and validity of its valuation system*. Melbourne: National Centre for Health Program Evaluation; 1995: Working Paper 42.
9. Sintonen H, Pekurinen M. A fifteen-dimensional measure of health-related quality of life (15D) and its applications. In: Walker S, Rosser R, eds. *Quality of life assessment*. Dordrecht: Kluwer Academic Publishers; 1993.
10. Sintonen H. *The 15D measure of health-related quality of life: reliability, validity and sensitivity of its health state descriptive system*. Melbourne: National Centre for Health Program Evaluation; 1994: Working Paper 41.
11. EuroQol Group. EuroQol: a new facility for measurement of health-related quality of life. *Health Policy* 1990; 16: 199-208.
12. Kind P. The EuroQol instrument: an index of health-related quality of life. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*. 2nd edn. Philadelphia, PA: Lippincott-Raven Publishers; 1996.
13. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. *J Clin Epidemiol* 1998; 51: 1115-28.
14. Murray C, Lopez A. *The global burden of disease*. Geneva: World Health Organization; 1996.
15. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) Instrument: a psychometric measure of health related quality of life. *Qual Life Res* 1999; 8: 209-24.
16. Hawthorne G, Richardson J, Day N, Osborne R, McNeil H. *Construction and utility scaling of the Assessment of Quality of Life (AQoL) instrument*. Melbourne: Centre for Health Program Evaluation; 2000: Working Paper 101.
17. Hawthorne G, Richardson J. *An Australian MAUI/QALY instrument: rationale and preliminary results*. Melbourne: Centre for Health Program Evaluation; 1995: Working Paper 49.
18. WHO. *International classification of impairments, disabilities and handicaps*. Geneva: World Health Organization; 1980.
19. Pedhazur E, Schmelkin L. *Measurement, design and analysis: an integrated approach*. Hillsdale: Lawrence Erlbaum; 1991.
20. Streiner D, Norman G. *Health measurement scales: a practical guide to their development and use*. 2nd edn. Oxford: Oxford Medical Publications; 1995.
21. Richardson J, Hawthorne G. *Negative utility scores and evaluating the AQoL: all worst health state*. Melbourne: Centre for Health Program Evaluation; 2000: Working Paper 73.
22. von Winterfeldt D, Edwards W. *Decision analysis and behavioural research*. Cambridge: Cambridge University Press; 1986.
23. Hawthorne G, Osborne R, McNeil H, Richardson J. *The Australian multi-attribute utility (AMAU): construction and initial evaluation*. Melbourne: Centre for Health Program Evaluation; 1996: Working Paper 53.
24. Segal I, Day N, Day S, Dunt D, Piterman H, Robertson I, et al. *Evaluation of the southern health care network co-ordinated care trial: final report: executive summary*. Melbourne: Centre for Health Program Evaluation; 2000.
25. Hogan A, Hawthorne G, Kethel L, et al. Health-related quality of life outcomes from adult cochlear implantation: a cross-sectional study. *Cochlear Implants Int* 2001 (in press).
26. Ware J, Snow K, Kosinski M, Gandek B. *SF-36 Health survey: manual and interpretation guide*. Boston: The Health Institute, New England Medical Centre; 1993.
27. WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychol Med* 1998; 28: 551-8.
28. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996; 15: 209-31.
29. Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994; 39: 7-21.