MONASH University
Business and Economics

Centre for Health Economics

Research Paper 2011 (71)

Psychometric Validity and the AQoL-8D Multi Attribute Utility Instrument

Jeff Richardson

Foundation Director, Centre for Health Economics
Monash University

Angelo Iezzi

Research Fellow, Centre for Health Economics
Monash University

October 2011

Centre for Health Economics
ISSN 1833-1173
ISBN 1 921187 70 0

Correspondence:

Professor Jeff Richardson
Centre for Health Economics
Faculty of Business and Economics
Monash University   Vic   3800
Australia

Ph: +61 39905 0754   Fax: +61 3 9905 8344
Email: Jeffrey.Richardson@monash.edu

ABSTRACT

The concept of validity and the phrase ‘an instrument has been validated’ are widely misunderstood. This has resulted in greater confidence in instruments than is justified. This has been used to bestow greater authority upon instruments than is justified by the empirical data. The present paper reviews the psychometric concept of validity as a prelude to its main focus which is the content and economic validity of a new multi attribute utility (MAU) instrument, the AQoL-8D. The AQoL-8D is a 35 item instrument with 8 dimensions (independent living, pain, senses, relationships, mental health, coping, happiness and self worth) which may also be combined into two ‘super dimensions’ for mental and physical health. The paper describes the stages of the analysis which led to the adoption of these dimensions and the methods used for achieving validity. It concludes by reporting several tests of the instrument’s validity.

TABLE OF CONTENTS

1 Introduction 1
2 The Problem 2
3 Validity 6
4 Constructing AQoL 8D 10
5 Tests of content validity 16
6 Conclusions 20
References 21

List of Tables

Table 1 Proportion of variance in one instrument explained by another instrument (R^2): Australia and USA 2
Table 2 Comparison of the dimensions and content of 6 MAU instruments 5
Table 3 Test-Retest reliability: intra class correlation[a] coefficients (ICC) 13
Table 4 Intra class correlation (ICC) between 8 instruments 17

List of Figures

Figure 1 Pair-wise comparison of 4 MAU instruments 4
Figure 2 Concept and item overlap 7
Figure 3 Construction of the descriptive system 11
Figure 4 The AQoL-8D model (35 items) 12
Figure 5 Mean analysis: actual TTO against predicted TTO 14
Figure 6 Frequency distribution of AQoL-8D utilities for the general population 14
Figure 7 AQoL-8D and dimension scores by age group, general population n = 884 15
Figure 8 Convergent validity: Comparison of AQoL-8D dimension scores for (a) general respondents; (b) mental health patients; and (c) patients awaiting bariatric surgery 16
Figure 9 Average Personal Wellbeing Index (PWI) score – total AQoL-8D 19

List of Boxes

Box 1 Tests of Validity 8
Box 2 AQoL instruments 10

Psychometric validity and the AQoL-8D multi attribute utility instrument

1 Introduction

Health economic evaluation increasingly uses cost utility analysis (CUA) in which benefits are (largely) measured by Quality Adjusted Life Years (QALYs). QALYs, in turn, are calculated by multiplying life years by an index of utility: the strength of preference for the health state of the life years. Utility may be measured in health state specific studies but, more commonly, it is calculated using a multi-attribute utility (MAU) instrument. This consists of two parts: a set of questions, the subject matter of which defines the instrument's 'descriptive system' and a scoring algorithm which converts answers into utility scores.

The objective of the Assessment of Quality of Life (AQoL) program was to create MAU instruments which achieved a high level of 'content validity'. This is a necessary but not sufficient condition for economic validity, ie for the measurement of what is conceptually needed for CUA. This raises the question why the program was necessary when a number of extant MAU instruments have been extensively validated. The answer, discussed in Section 2, is that the process of 'validation' commonly reported in the literature does not ensure either content or economic validity. Confidence in the instruments appears to be a result of the compelling connotations of the term 'validated' rather than a reflection of empirical evidence.

In broad terms 'validation' is a *process* for increasing confidence that an instrument will give correct prediction (Streiner and Norman 2003). This is achieved a number of ways as summarised in Section 3. Two types of 'validity' are of particular interest. In the context of QALYs, 'economic validity' means that the utility score will correctly predict the trade-off a person would select between the quantity and quality of life. Related to this, 'content validity' is the requirement that the utility score fully takes account of all of the elements in a health state that are relevant to a person's preferences.

The approach adopted for achieving these two properties in the AQoL-8D instrument is described in Section 4. Content validity was sought through the use of psychometric procedures for construction of the instrument's descriptive system. Economic validity was sought by the adoption of a two stage methodology for deriving the scoring algorithm (utility formula). Some tests of the content validity of the instrument are reported in Section 5.

2 The Problem

Existing MAU instruments do not predict the same utility scores and, in view of the fact that they purport to measure an identical variable (utility), the correlation between them is low.

There have been surprisingly few multi-instrument comparisons. (For a review see Richardson, McKie, Bariola (2011c)). In an early Australian comparison 956 hospital and general respondents were administered the EQ-5D, SF-6D, 15D, HUI 3 and AQoL-4D. Table 1 reports the proportion of the variance in each instrument's scores explained by each of the other instruments. Overall an average of 44 percent of the variance is unexplained. In the more recent US study (Fryback et al. 2010), 3844 adults were surveyed to compare the EQ-5D, QWB^{SA}, HUI 3 and SF-6D . A weaker association was found than in Australia. An average of 53 percent of these instruments' variance was not explained by the other instruments. Generally researchers conducting multi-instrument comparisons have concluded that the utilities derived from them are 'not equivalent', that translation between them will result in 'low precision' and that comparisons between them 'warrant caution'.

Table 1 Proportion of variance in one instrument explained by another instrument (R^2): Australia and USA

7A Australia	15D	EQ5D	HUI 3	SF-6D	AQoL-4D
15D	1.00				
EQ5D	0.58	1.00			
HUI 3	0.55	0.41	1.00		
SF6D	0.59	0.56	0.44	1.00	
AQoL	0.64	0.53	0.55	0.55	1.00
MEAN	0.59	0.52	0.49	0.53	0.57
7B USA	**QWB SA**	**EQ5D**	**HUI 3**	**SF6D**	
QWB SA	1.00	0.41	0.45		
EQ5D		1.00			
HUI 3		0.49	1.00		
SF6D	0.43	0.50	0.52	1.00	
MEAN	0.43	0.47	0.49	0.48	

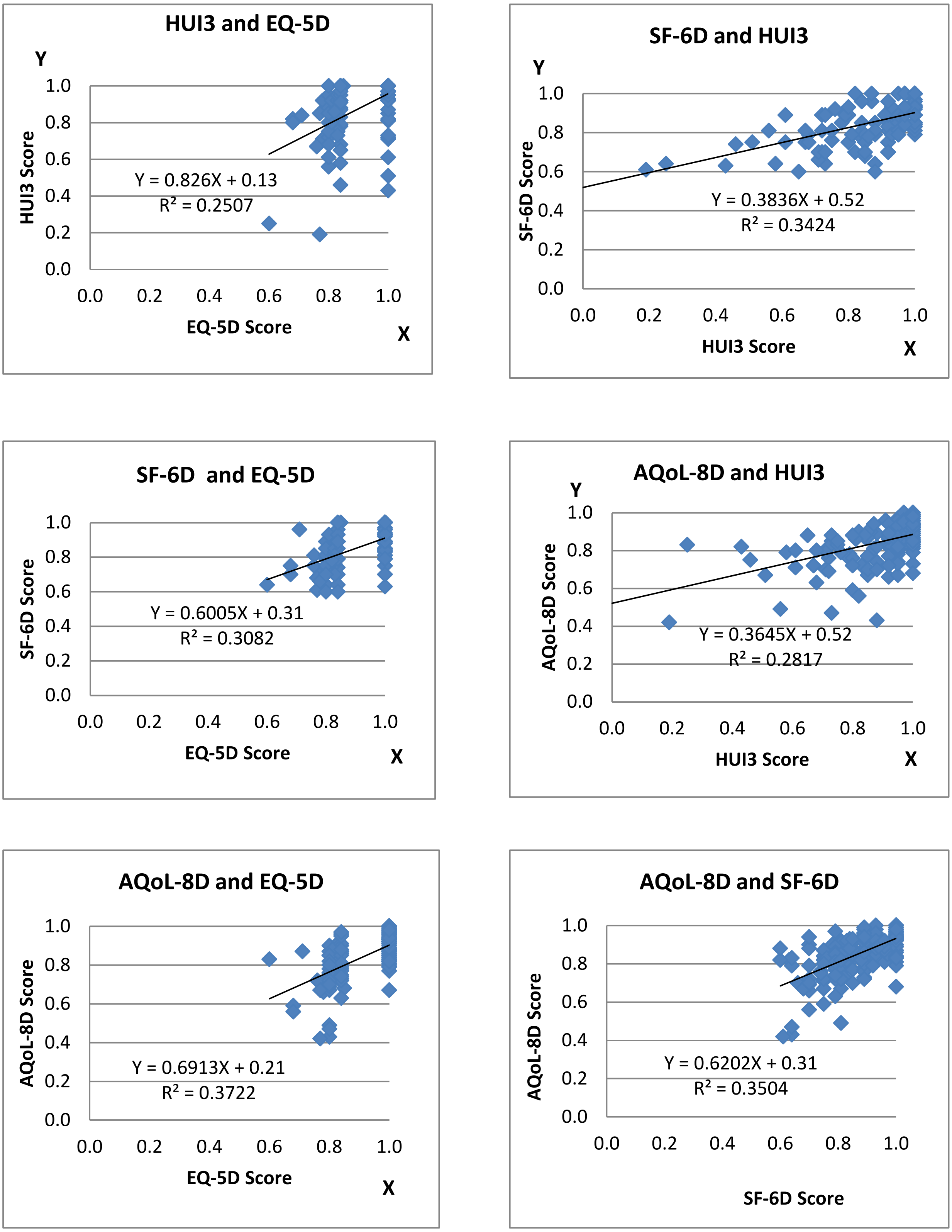
Source: Hawthorne & Richardson (2001); Fryback, Palta et al. (2010).

The low correlation between instruments is necessarily attributable to either the scoring formula or the content of the descriptive systems. However, few comparative studies have considered how differences in content affects instrument validity. The limited evidence indicates that the effects are large. The simplest test is a comparison of upper end sensitivity (ceiling effects) as indicated by the percentage of respondents who are estimated to have the maximum or close to maximum possible score. In the early Australian study significant differences were found. In the US study the percentages of scores above 0.95 were 37.0 (EQ-5D); 36.9 (HUI 2); 36.2 (HUI 3); 1.7 (SF-6D) and 2.3 (QWB). Figure 1 illustrates the differences and, more generally, the low correlation between instruments using results from a recent survey of Melbourne's Bangladeshi population (Khan and Richardson 2009). The data reflect the strong ceiling effect of the EQ-5D (the horizontal scale in the three left hand diagrams) and the significant, but weaker ceiling effect of the HUI 3. The SF-6D and EQ-5D have the strongest floor effect(s) with no values below 0.6. The AQoL-8D and the HUI had minimum scores of 0.42 and -0.04 respectively.

More generally the scattergrams in Figure 1 illustrate the lack of correspondence between predicted utilities. The line of best fit between instrument scores might be used to abstract from individual differences. However in each case this would lead to very different prediction with respect to the effect of change. For a similar prediction the 'b' coefficient in the regression equation would equal 1.00. As shown, the coefficient varies from 0.36 to 0.82. In the present context, however, the problem of variable instrument content is best illustrated by differences in scores from the same individual completing the different instruments. At all levels of one instrument there was significant variation in the value of other instruments. When SF-6D = 0.6, HUI 3 and AQoL-8D values varied from (0.2-0.9) and (0.4-0.9) respectively; when AQoL-8D = 0.8 HUI 3 and SF-6D varied from (0.25-1.00) and (0.60-1.00) respectively. Importantly, differing results were obtained from the same individuals and the magnitude of the discrepancies in the utility scores to be explained is indicated by the extreme range of individual differences and not by average differences in group utility scores. Some of this variation would be random. A small amount can be attributed to the choice of preference instrument and scoring model but a large amount must be attributed to the instrument descriptive systems.

These are contrasted in Table 2 which includes the SF36 as a reference. The six instruments display enormous variation with imperfect correspondence between the dimensions included and the number of items per dimensions. Item descriptions vary and item response levels vary from 3 to 6, resulting in 243 health states for the EQ-5D, 3.1 x 10^{10} health states for the 15D and 2.37 x 10^{23} health states for the AQoL-8D. Differences are so great that differences in content validity and estimated utility scores should be unsurprising.

Figure 1 Pair-wise comparison of 4 MAU instruments



Source: Khan and Richardson (2011)

Table 2 Comparison of the dimensions and content of 6 MAU instruments

		Number of symptoms (.) and items (*)					
	Dimension	SF36	15D[2]	EQ-5D	HUI 3	SF-6D (36)	AQoL-8D
Physical	Physical ability/ vitality/Coping/ Control	***	*			*	*
	Bodily Function/ Self Care	*	***	*			*
	Dexterity	*			*		
	Pain/Discomfort	**	*	*	*	*	**
	Senses		**		**		**
	Usual activities/ Work function	****	*	*	*	*	****
	Mobility/walking	*****	*	*	*		*
	Communication		*		*		*
	Sickness, health	****					**
Psycho-social	Sleeping	*	*				*
	Vitality, emotions	***					*
	Tranquillity	*	*				**
	Psychological: Depression/Anxiety/ Anger	****	*	*	*	*	******
	General Satisfaction, happiness	**					***
	Self Esteem						**
	Cognition/Memory Ability		*		*		
	Social Function/ Relationships	***				*	****
	(Family) Role					*	*
	Intimacy/Sexual Relationships		*				*
			15 items	5 items	8 items	6 items	35 items

Notes:

1 Symptom problem groups associated with consciousness, burns, pain, stomach, cough, fever, depression, headache, itching, talking, eyes, weight, teeth, ears, hearing, throat, breathing, sleeping, intoxication, sex, anxiety, eyeglasses, use of medication.

2 15D also includes breathing, sleeping, eating, elimination, sexual activity.

3 Validity

In principle the concept of validity is straightforward. A valid instrument measures what it purports to measure. A correctly calibrated ruler, for example, gives accurate measurement of distance. Validation of constructs such as intelligence or the quality of life however is more complex. There is no ‘gold standard’ as there is for physical measurement. A construct such as intelligence is commonly the result of a number of elements: verbal, numerical, spatial skills, problem solving, memory, etc. In turn, each of these may not be clearly identified by the answer to a simple question, but may require a series of questions and answers. Further, the precise meaning of terms and questions can vary between individuals and cultures in a way which is related to personal circumstances. As an example, ‘communication’ may mean speaking to some, signing to others, face to face contact for some, or texting for others. Happiness may primarily be dependent upon social relationships in a culture which is community oriented (Asia) but self-referential in individualistic (western) culture.

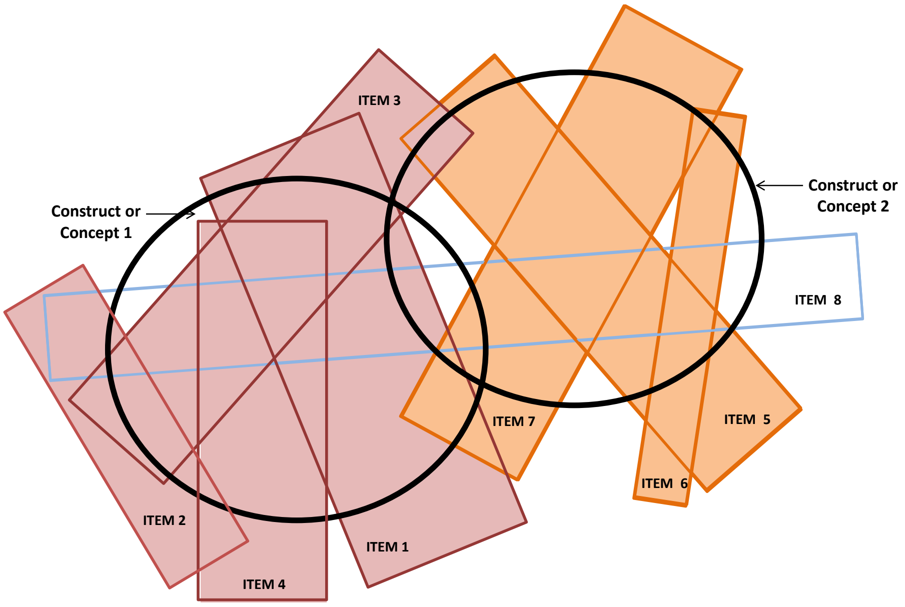
To overcome this problem psychometric ‘Classical Test Theory’ uses different factor analyses to create measurement instruments.[1] Answers to questions are analysed to determine their relationship, and answers which cluster around a concept – the answers correlate – are accepted as a measure of this concept. This is illustrated in Figure 2 in which two constructs or concepts are represented in ‘content space’ by the heavy bold circles. A series of questions and answers – items – are represented by the various rectangles. As shown, four of these heavily overlap Concept 1. Three overlap Concept 2. Item 8 crosses both concepts. In the terminology of factor analysis this last item ‘cross loads’ on the two concepts and would normally be eliminated from the items used in an instrument.

Figure 2 illustrates a number of points. Statements overlap and do not exactly correspond with concepts. Importantly, single statements may cover only a small part of the content of a concept – ie language and concepts are imperfectly related. Finally, as shown, neither concept may be perfectly defined by the items. Some content may be omitted by the item description.

In practice validation is a process of hypothesis testing: increasing the confidence we have in an instrument and confidence in the inferences drawn from it. This implies that an instrument can never be fully validated: we have more or less confidence in it. Importantly, the more demanding the test the greater the confidence. The less demanding the test the less the confidence.

[1] It is necessary to distinguish CTT (Classical Test Theory) from IRT (Item Response Theory). CTT has recently used factor analyses to create measurement instruments. Historically, the foundation concepts (and strategies) for CTT were the item-total (or the item-remainder) correlations (that provided evidence of item discrimination) and a measure of internal consistency (eg Cronbach’s alpha). In contrast, the ‘foundational’ measurement concept for IRT is the item response curve. IRT seeks items that relate to one single latent trait that satisfy the criterion of ‘conditional independence’ – no association between items over and above that explained by the one latent variable (Fayers and Machin 2000).

Figure 2 Concept and item overlap



Key:
'Item' = question with a series of possible response levels (eg How often do you feel sad? a) never, b) rarely, c) some of the time, d) usually, e) nearly all the time)
Concept = an abstract idea concerning some hypothesised attribute or characteristic (physical fitness, mental health)
Construct = A mini theory or created construct to explain observed behaviour.

Source: Richardson (2010)

The different tests of validity have been variously labelled, as described in Box 1. In the absence of a gold standard, validation of a construct (construct validity) usually refers to content, concurrent or predictive validity. The relationship between them is described by Streiner and Norman (2003):

> "A measure that includes a more representative sample of the target behaviour lends itself to more accurate inferences; that is inferences which hold true under a wider range of circumstances. If there are important aspects of the outcome that are missed by the scale, then we are likely to make some inferences which will prove to be wrong; our inferences (not the instrument) are invalid" (page 175).

Achieving a valid instrument for constructing QALYs is challenging because of the multiple and conflicting requirements for validity. These include:

- Sensitivity (construct validity)
- Non redundancy (economic validity)
- Formative and reflective modelling (content validity)
- Use of a preference based scaling instrument (economic validity)

Box 1 Tests of Validity

Different tests have been described under different headings. The common endpoint, however, is increased confidence in the inferences made from instrument scores.

Classification of tests:

- Translation or representation validity
  - Face validity
  - Content validity
- Construct validity
  - Convergent validity
  - Discriminant validity
- Criterion validity
  - Concurrent validity
  - Predictive validity

Translation or representation validity: A general term for the extent to which a construct (concept) can be successfully translated into, or represented by, specific tests.

Face validity: The instrument seems, at face value, to capture the construct, for example, by naming it. This is generally considered the weakest form of test

Content validity: The extent to which an instrument includes or covers a representative sample of the construct's behaviour domain, for example, determining arithmetic skill by asking for the answers to 3-4 questions for each domain of arithmetic – addition, subtraction, multiplication, division, fractions, decimals, etc.

Criterion validity: A general term for the use of some external criterion to test the concept.

Concurrent validity: An instrument can distinguish, as expected, between groups, for example, the general population and hospital patients.

Predictive validity: The ability to predict what is expected. This includes the predictive tests above but is more general. For example an IQ test may predict subsequent income.

Construct validity: A general term for the success of a test or instrument in measuring a construct (concept). It commonly subsumes convergent and discriminant validity.

Convergent validity: A specific test of construct or criterion validity. Instrument scores correlate, as predicted, with other instrument scores or some criterion score which are known to correlate with the construct. For example the HUI 3 instrument correlates with the EQ5D.

Discriminant validity: A specific test of content or criterion validity. Instrument scores do not correlate with instrument scores unrelated to the construct. For example EQ5D scores would be expected to have low correlation with a person's blood pressure.

First, and for the reasons outlined above, the instrument is likely to require multiple overlapping items to achieve content validity (sensitivity). Secondly, however, overlapping items result in the double counting of some elements. For this reason decision analytic theory requires that items be orthogonal to avoid redundancy (double counting). This is usually achievable in the context in which decision theory evolved. The problem of where to locate a car factory can be broken down into a number of independent criteria (items): the cost of labour, capital, location, colour, shape and power of the engine, etc. The items do not overlap. This generally cannot be achieved with psychometric constructs. Consequently a method must be devised for reconciling the sensitivity of content with the resulting redundancy.

This problem only exists if redundancy matters. It is commonly ignored in psychometric instruments and the interpretation of scores is simply adjusted. Excellent health, as measured by the SF36, is ‘norm referenced’. The corresponding score depends upon the number of questions (36) the number of response categories and the responses obtained from those with excellent health. In contrast, the numbers on a utility scale suitable for QALY measurement are ‘criterion referenced’ with the length of life as the criterion. For ‘economic validity’, a 10 percent reduction in the index of utility must have the same effect upon preferences as a 10 percent reduction in the quantity of life. The most demanding (and neglected) test of the validity of utility scores is independent evidence that the implied trade-off between the length and quality of life is acceptable.

A final complexity with the construction of the MAU descriptive system arises from the distinction between formative and reflective modelling. With reflective modelling, causation runs from the latent (unobserved) variable(s) to the items. (For example, it might be hypothesised that a latent variable for ‘intelligence’ is causally related to observed items measuring arithmetic and linguistic achievements, as intelligence *causes* the high achievement measured by the items.) With formative models causation is reversed. For example, socio economic status (SES) does not cause education or income; rather education and income, *inter alia*, define SES.

The descriptive system of a sensitive MAU instrument has elements of both types of model. As discussed, dimensions of health are reflective. Depression causes people to feel sad, despair, sleep badly and self harm, not vice versa: the depression causes the symptoms and could be identified with different combinations of symptoms. However, there are formative elements. There are a number of broad concepts which are not symptoms of the quality of life, but are defining characteristics. Omission of pain, mobility and depression (or proxy items for them) would violate the usual concept of quality of life and an instrument which omitted them would lack content validity.

Finally, for economic validity, utility measurement must be based upon a scaling instrument which is accepted as measuring the strength of preferences. While this topic is large and controversial it is not discussed here as it has been the (almost exclusive) focus of discussion of validity in the literature.

4 Constructing AQoL 8D

The AQoL instruments are described in Box 2. Each was a result of two analyses which resulted in the construction of the instrument's descriptive system (survey questionnaire) and the instrument's utility scoring formula (algorithm).

Box 2 AQoL instruments

AQoL-4D Originally called 'AQoL' (Hawthorne, Richardson, and Osborne 1999):
Initially a 5 dimension 15 item instrument. Dimensions were illness, independent living, social relationships, physical senses, psychological wellbeing. Illness was subsequently deleted. Utilities were combined with a multi-level model using multiplicative models for dimensions and an overall multiplicative model to combine them.

AQoL 8 (Hawthorne 2009)
An 8 item (Brief) instrument which removes one item per dimension from AQoL-4D.

AQoL-6D (Richardson et al. 2004):
A 6 dimensional 20 item instrument. Pain and coping were added to AQoL-4D as separate dimensions. Mental health and independent living items were increased from 3 to 4. Utility weights were constructed as for AQoL-4D but with an econometric adjustment for the final algorithm.

AQoL-7D (Misajon et al. 2005; Richardson et al. 2011b):
A 7 dimension 26 item instrument which adds an explicit dimension for vision (VisQoL). Scaling was carried out as for AQoL-6D

AQoL-8D (Richardson et al. 2011a; Richardson et al. 2011d):
The 8 dimensional 35 item instrument shown in this paper.

4.1 Descriptive System: The steps which led to the descriptive system are outlined in Figure 3 and follow those recommended by instrument construction theory (Fayers and Machin 2000; Streiner and Norman 2003). Details are provided in Richardson, Elsworth et al. (2011a).

Theory: The first step is to determine the overall theory of HRQoL to be embodied in an instrument. This determines the type of dimensions and items to be included in an instrument.

Items and content analysis: A total of 250 items were compiled from a review of psychological instruments, from four focus groups and from the research team which included two psychiatrists, a psychologist, a counsellor, a psychometrician and a health economist. Items were triaged to eliminate obvious redundancy, poorly worded and ambiguous items. The inclusion of the items from the AQoL-4D, AQoL-6D and the K10 psychological distress scale resulted in a questionnaire of 135 items.

Construction survey: The 135 items were administered to 716 respondents: 514 mental health patients who were interviewed and 202 members of the public who retuned useable mail questionnaires, a response rate of 40.6 percent of eligible contacts.

Statistical analysis: Item and dimension selection were conducted in the tradition of Classical Test Theory using a combination of unrestricted and restricted (exploratory and confirmatory) factor analyses. This permitted the unrestricted exploration of item combinations to produce dimensions with psychometric integrity (all items 'loaded' upon a single latent variable which contributed significantly to the dimension content) while simultaneously constraining the overall instrument to contain theoretically required dimensions. In practice, this meant the iterative exploration of combinations of items and dimensions and their testing to determine whether they constituted a well-constructed instrument as indicated by its diagnostic statistics.

An initial 32 item instrument, PsyQoL, was constructed from seven psycho-social dimensions. These were reduced to a PsyQoL-Bref instrument and combined with the AQoL-6D to form the AQoL-8D. PsyQoL and PsyQoL-Bref are reproduced in Richardson Elsworth et al. (2011) and appear on the AQoL website [http://www.aqol.com.au/].

The AQoL-8D is shown in Figure 4. Along with other AQoL instruments it is unique amongst MAU instruments in its hierarchical structure. This helped reconcile the need for reflective and formative modelling. The dimension content reflects the dimension latent variables which could be investigated with unrestricted analyses. The overall instrument structure and choice of dimensions, however, was restricted for theoretical reasons as discussed above.

Figure 3 Construction of the descriptive system

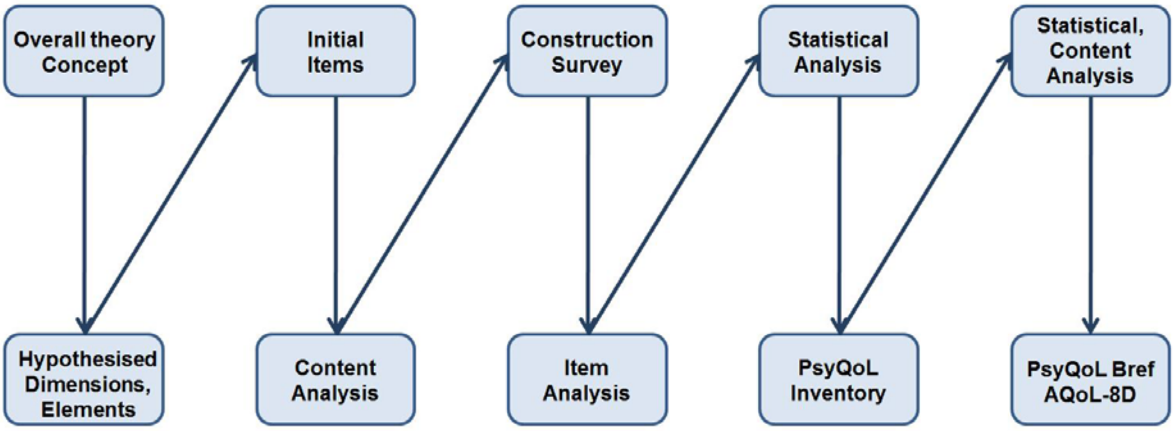
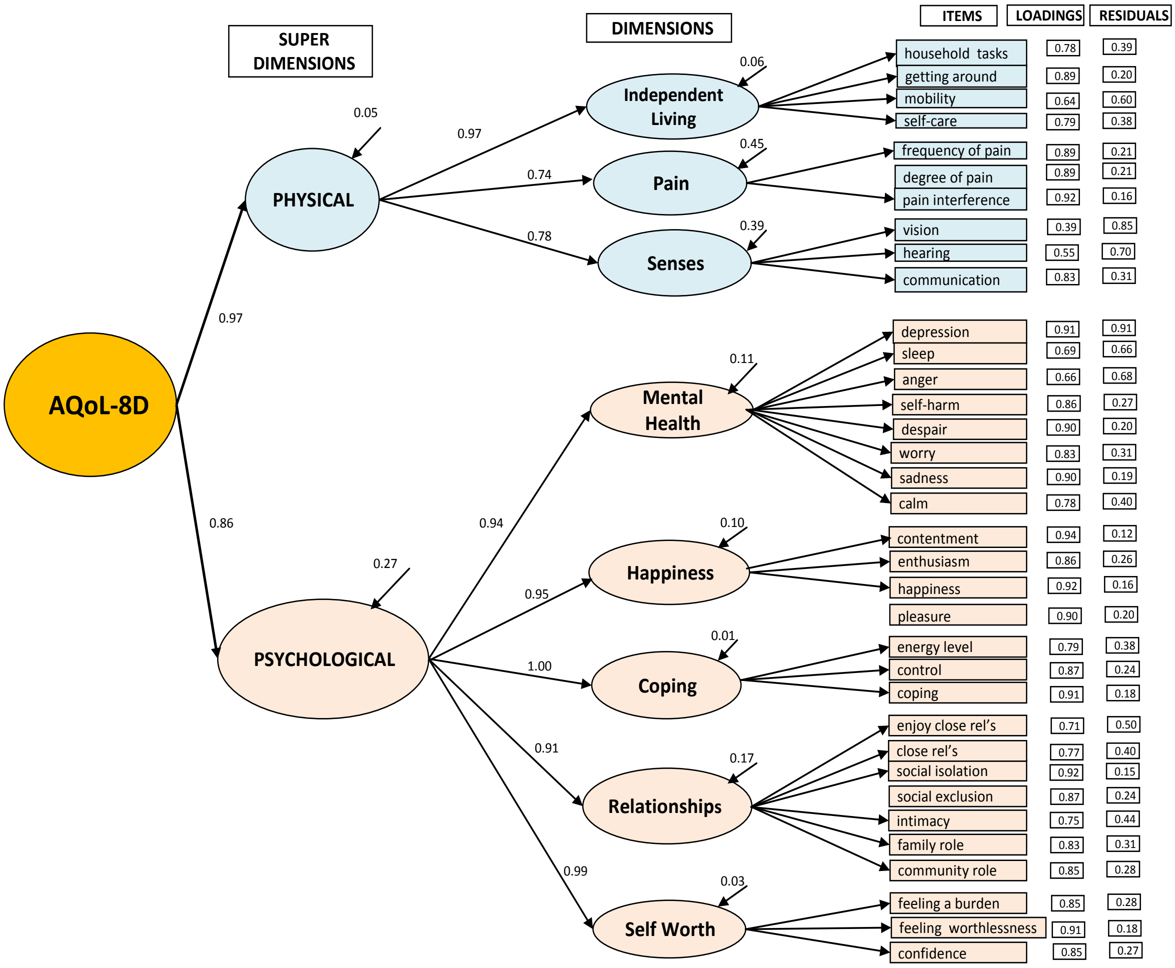


Figure 4 The AQoL-8D model (35 items)



Fit Statistics: CFI = 0.974; TLI = 0.972; RMSEA = 0.073; WRMR = 1.64

Notes:

1. Numbers pointing to constructs are the residual (unexplained) variances of the latent variables (also called disturbances).
2. Numbers on arrows between constructs are factor loadings.
3. Unstandardised loadings of AQoL-8D on Physical and Psychological were constrained to be equal identification.

4.2 Utility formula: The derivation of the utility formula is described in Richardson, Sinha, Iezzi et al. (2011d). This drew upon a survey of 670 individuals, 323 patients and 347 members of the public. All respondents completed an initial questionnaire and were subsequently interviewed. This provided data for a three part modelling of the instrument shown in Figure 4.

Part 1 Multiplicative dimension modelling: Item responses were weighted and combined using a multiplicative model (similar to HUI 1-3 and AQoL-4D, 6D).

Part 2: Dimension scores were weighed and combined also with a multiplicative model.

Part 3: Regression analyses were used to correct the multiplicative model and to eliminate the effects of content redundancy caused by item and dimension overlap. Independently collected TTO scores from multi attribute health states were regressed upon the multiplicative scores for both the overall AQoL-8D and each of the dimensions. The final algorithm for dimensions, super dimensions and the overall model are on the AQoL website.

The need for the three part procedure arose from the constraints outlined earlier. Content validity usually requires a minimum of 3-4 items per dimension. This results in both content overlap (redundancy) and an instrument whose size makes the use of one stage econometric methods problematical. Multiplicative modelling is unconstrained by the number of dimensions but cannot be an endpoint as it presupposes orthogonality – non-redundancy. The combination of multiplicative and econometric modelling used for AQoL-8D overcomes the problem. Stages 1 and 2 reduce the analysis by collapsing the number of independent variables from 35 (items) to 9 (multiplicative scores for AQoL and each of the 8 dimensions). Econometric results may then be used to extrapolate and interpolate utility scores for all other health states.

4.3 Properties of the AQoL-8D: The final relationship between the independently assessed health states and those predicted by AQoL-8D are shown in Figure 5. From the resulting frequency distribution (Figure 6) there are neither significant ceiling nor floor effects and from Table 3 there is good test-retest reliability.

The age profile differs by dimension. The physical dimension scores decline while mental health, relationships, self worth and happiness improve. The overall AQoL-8D utility increases insignificantly. The dimension scores, shown in Figure 7, cannot be directly compared as dimension scales (means, standard deviations) differ.

Table 3 Test-Retest reliability: intra class correlation[a] coefficients (ICC)

	Happiness	Mental Health	Coping	Relation-ships	Self-worth	Ind Living	Pain	Senses	Super dimensions		AQoL-8D
									Mental	Physical	
Base – 2 weeks	.858	.870	.816	.783	.863	.861	.851	.644	.902	.842	.907
Base – 1 month	.846	.844	.795	.733	.848	.856	.851	.691	.863	.874	.894

Source: Richardson, Sinha, Iezzi et al. (2011d)

Figure 5 Mean analysis: actual TTO against predicted TTO

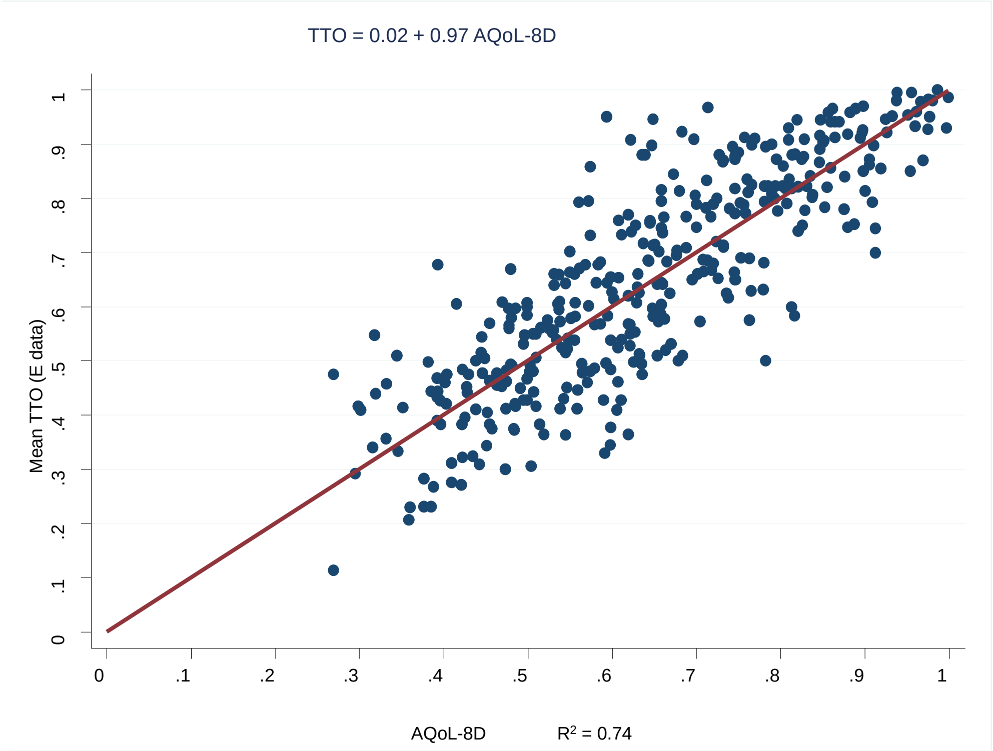


Figure 6 Frequency distribution of AQoL-8D utilities for the general population

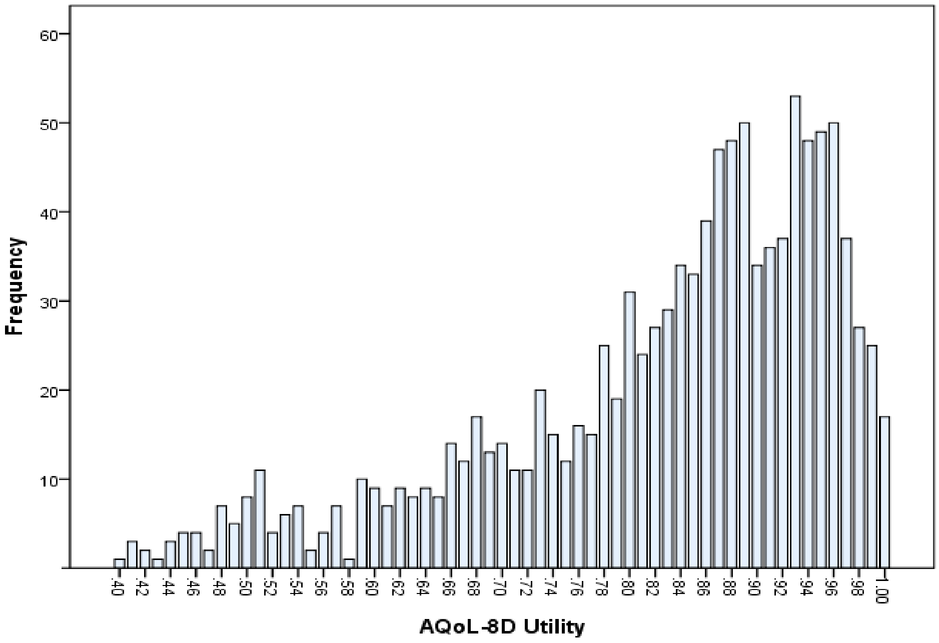
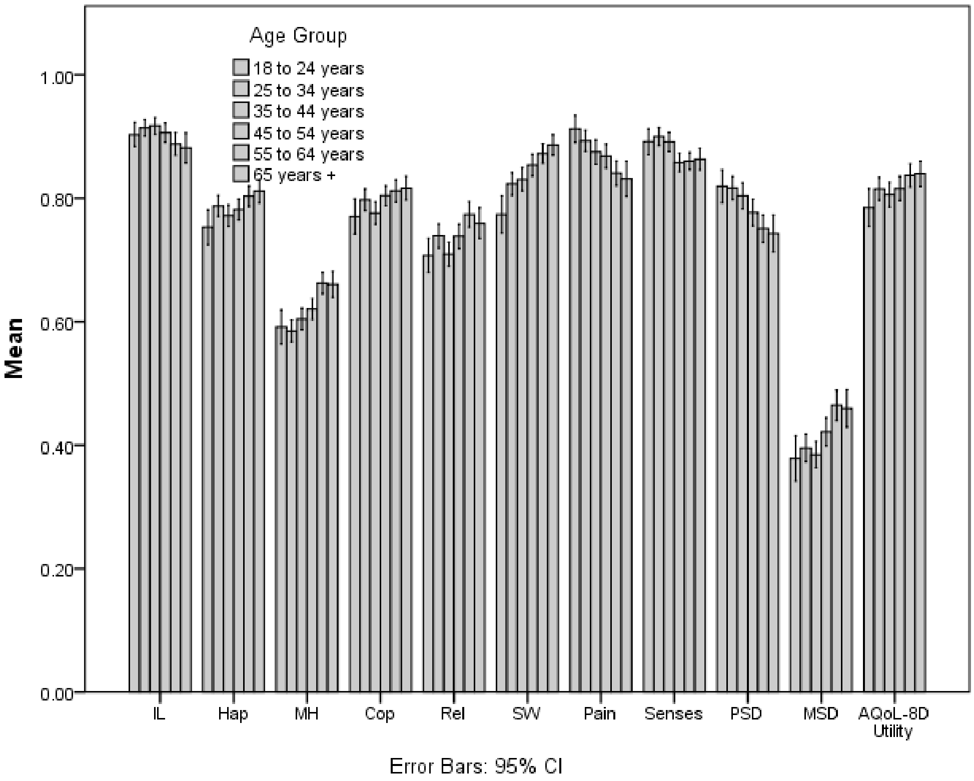


Figure 7 AQoL-8D and dimension scores by age group, general population n = 884



5 Tests of content validity

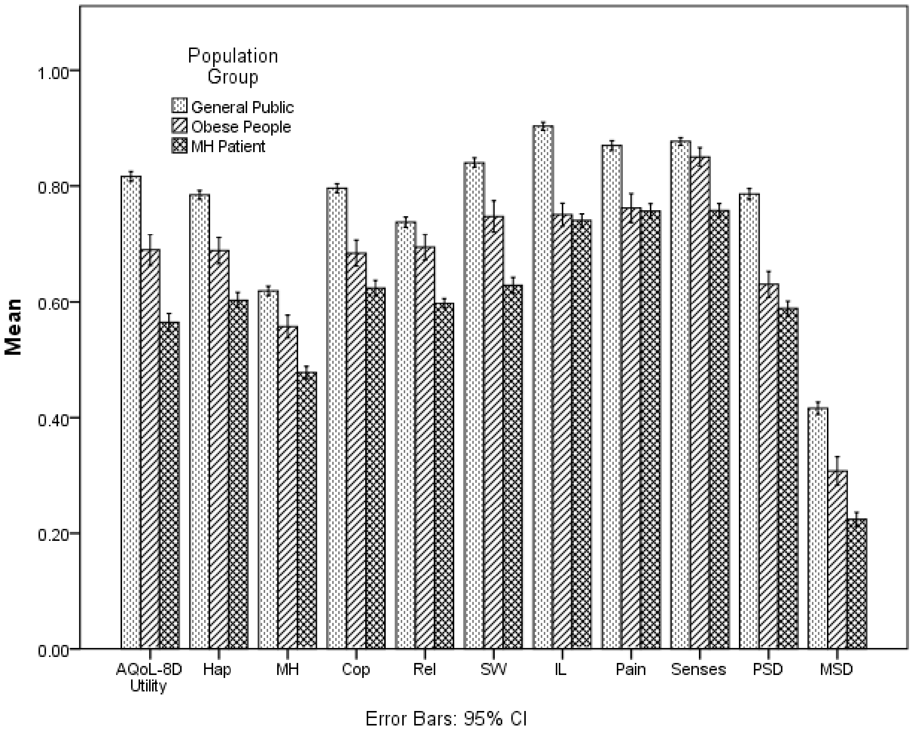
As outlined, validation is a process of hypothesis testing to build confidence in prediction and this may be achieved a number of ways.

Face and content validity: These tests of validity have been discussed in the earlier sections of the paper. Comparison of items and dimensions with the content of other widely accepted instruments provides – arguably the weakest – evidence for validity. Table 2 indicates that there is significant correspondence between the content of AQoL-8D and SF36 (a non-utility instrument) and, with the exception of dexterity and cognition AQoL-8D has items in each of the dimensions described by other MAU instruments.

Greater support for content validity is obtained from the construction process, ie from the extent to which the instrument's satisfactory description of the content of the items in the original item base from which it was derived and from the psychometric properties of the final instrument. AQoL instruments were unique amongst MAU instruments in seeking this form of validation and the diagnostic statistics reproduced in Figure 4 indicate that AQoL-8D performs well in this respect.

Convergent validity: Instruments should discriminate between populations in different health states and in the expected way. Figure 8 compares dimension and overall utility scores for members of the general public, mental health patients and patients awaiting bariatric surgery. Utilities differ significantly and as expected by dimension.

Figure 8 Convergent validity: Comparison of AQoL-8D dimension scores for (a) general respondents; (b) mental health patients; and (c) patients awaiting bariatric surgery



Notes: n (general population) = 884; n (obese patients) = 196; n (mental health patients) = 832

Predictive validity: Instruments should predict what is expected. This is commonly tested using the correlation between instruments. Results of a recent multi-instrument study including the AQoL-8D are reported in Khan and Richardson (2011). The unique feature of this study of the Melbourne Bangladeshi population is that, in addition to including four MAU instruments, it included two subjective wellbeing instruments, a psychological distress and an overall self-assessment scale. Correlation between these is shown in Table 4. The average correlation between the MAU instruments is low (partly due to the target population having few unhealthy members) and there is little difference between their predictive validity using this test.

Table 4 Intra class correlation (ICC) between 8 instruments

Measures	Correlations							Highest correlation with:
	1	2	3	4	5	6	7	
1. EQ-5D	1							AQoL-8D
2. HUI 3	.502**	1						SF-6D
3. SF6D	.558**	.586**	1					AQoL-8D
4. AQoL-8D	.610**	.531**	.593**	1				EQ-5D
Average (1-4)	**0.56**	**0.54**	**0.58**	**0.58**				
5. PWI	.452**	.521**	.476**	.496**	1			HUI 3
6. SWLS	.395**	.477**	.348**	.503**	.534**	1		AQoL-8D
7. K-10	.567**	.456**	.514**	.668**	.460**	.440**	1	AQoL-8D
Average (1-7)	**.51**	**.51**	**.51**	**.57**	**.49**	**.45**	**.52**	

** Correlation is significant at the 0.01 level (2-tailed).

Source: Khan and Richardson (2011)

However non MAU instruments correlate more highly with AQoL-8D than with other MAU instruments. The result is particularly interesting as it is the first time (to our knowledge) that subjective wellbeing (SWB) and MAU instruments have been compared as the usual presumption is that they tap into different cognitive domains (SWB into ‘affect’; MAU instruments into ‘cognition’). Given this presumption the relationship between Personal Wellbeing Index (PWI), a leading SWB instrument and AQoL-8D is striking. As shown in Figure 9 their relationship in this population is very close.

Economic validity: The most demanding validation test for an MAU instrument is that the utility score it produces embodies the defining property of a QALY, namely that a given percentage increase in the score is judged to be as desirable (there is the same preference for it) as the same percentage increase in the quantity of life. No definitive test of this property has been proposed to date. However the property may be subject to a test of predictive validity or what Nord described as a test of ‘reflective equilibrium’. Changes in utility can be used to calculate the changes in the quantities of life which are predicted to be of equal value. For example, it would be predicted that a lifelong increase in a person’s utility from 0.5 to 1.0 would be equally valued as a doubling in the person’s life expectancy (subject to time discounting). The equivalence of the value can, in principle, be judged independently.

Table 5 applies this logic to six instruments. Health states were selected which were as close to equivalent as the differing predictive systems permitted. The scoring algorithms for each was used to calculate the increase in utility which would occur if the health problem was removed and the person achieved the highest health state on the scale. The table reports the implication of these results for the quantity of life.

No test has been conducted to determine which of the instruments reflects preferences most accurately. However the information required to complete the test commenced in Table 5 is uncomplicated, at least, in principle. It requires judgement of the equivalence of the benefit shown in column 2 and the benefits shown in the final two columns. We postulate that, given the value normally attached to life per se, the benefit of curing mild to moderate pain for 20 years would be closer to the 3.5 month extension of life implied by AQoL than to the 9.6 year extension implied by the QWB or 11.1 year extension implied by the EQ-5D. A minimum conclusion which may be drawn from Table 5 is that there are major discrepancies between the instruments and that their economic validity – their ability to correctly predict the preferred trade-off between the quantity and quality of life – requires further investigation.

Table 5 Predictive validity: prediction from utility scores

Instrument	**Permanent problem cured**	**Increase in utility(1)**	**Equivalent**		
	= return to good health for 20 years	**Value p.a.**	**Cures = 1 life saved(2)**	**life extension with original QoL(3)**	
				RTP = 0%	**RTP = 2%**
QWB	Headache ***or*** dizziness ***or*** ringing in ears ***or*** spells of feeling hot, nervous ***or*** shaky	0.244	4	6.5 years	9.6 years
15D	Mild physical discomfort...pain, ache, nausea, itching, etc	0.023	4.3	5.6 months	8.3 months
EQ-5D	Moderate pain or discomfort, some problem walking	0.273	5	7.5 years	11.1 years
HUI 3	Moderate pain that prevents a few activities	0.137	7	3.2 years	4.7 years
SF-6D	Pain which interferes with normal work...a little bit	0.07	14	1.5 years	2.2 years
AQoL-8D	Moderate pain...which sometimes interferes with usual activities	0.01 (1)	100	2.4 months	3.5 months

Source: (Richardson, McKie, and Bariola 2011 forthcoming)

Notes

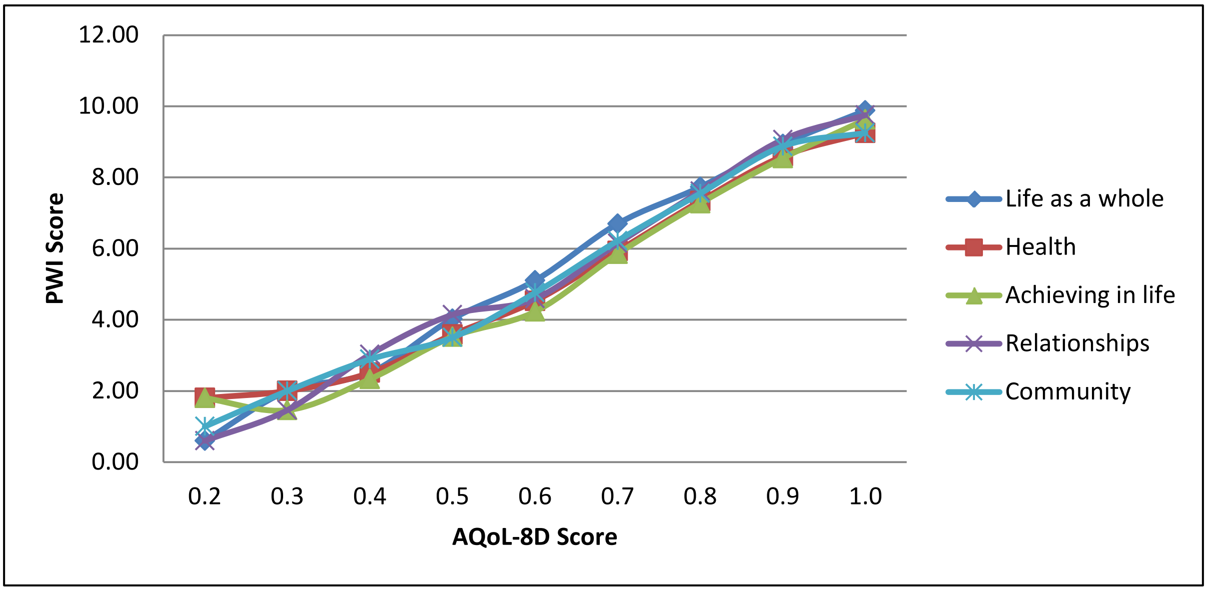
(1) Increase in utility if an individual is cured from the permanent problem and returned to normal or best health

(2) The number of cures, n, equivalent to saving one life is calculated as n = 1/(increase in utility). Therefore cures items value of cure = n x increase in utility = 1.00

(3) The number of years of life extension, n, is calculated from QALY gain = 20 (utility gain) = n.(original utility)

(4) AQoL-8D is at ‘normal’ (not best) levels for 7 additional items, viz, jobs around house, getting around the house, mobility, toileting, coping, relationships, content with life, enthusiasm

Figure 9 Average Personal Wellbeing Index (PWI) score – total AQoL-8D



6 Conclusions

A large number of studies have been conducted to validate the major MAU instruments. These have generally concluded that, to a greater or lesser extent, the scales are valid. Despite this, the utilities predicted differ significantly. It has been argued here that this is a reflection of the limited scope of the tests that have been used and a limited application of the process of 'validation'. The most common test – correlation of an instrument with another (criterion) instrument – is a necessary but not sufficient condition for validity. All instruments purporting to measure quality of life are likely to correlate and the test is not, consequently, a strong one.

It has been argued here that the differences in the predicted utilities are attributable, in large part, to the enormous differences in the instruments' descriptive systems and therefore their content validity. The subject has been largely ignored in the literature which has been primarily concerned with the techniques used for attributing utility scores to existing descriptive systems.

In contrast, the AQoL program focused upon content validity and the use of psychometric procedures for achieving this. As described, this task encounters a number of problems and, in particular, the trade-off between content sensitivity and redundancy and between the need for both formative and reflecting modelling. The most recently developed instrument – AQoL-8D – has strong psychometric properties which increase confidence in content validity and, additionally, it performs well in other validation tests. With respect to economic validity, the trade-off between the quantity and quality of life predicted by the AQoL-8D – at least in the vicinity of good health – differs significantly from other MAU instruments. This is an area requiring greater research.

References

Fayers, P. M. and D. Machin. 2000. *Quality of Life: Assessment, Analysis and Interpretation*. Chichester: John Wiley & Sons Ltd.

Fryback, D. G., M. Palta, D. Cherepanov, D. Bolt, and J. Kim. 2010. “Comparison of 5 health related quality of life indexes using item response theory analysis.” *Medical Decision Making* 30(1): 5-15.

Hawthorne, G. 2009. “Assessing utility where short measures are required: development of the short Assessment of Quality of Life 8 (AQoL 8) instrument.” *Value in Health* 12(6): 948-57.

Hawthorne, G., J. Richardson, and N. A. Day. 2001. “A comparison of the assessment of quality of life (AQoL) with four other generic utility instruments.” *Annals of Medicine* 33(5): 358-70.

Hawthorne, G., J. Richardson, and R. Osborne. 1999. “The Assessment of Quality of Life (AQoL) instrument: A psychometric measure of health related quality of life.” *Quality of Life Research* 8: 209-24.

Khan, M. A. and J. Richardson. 2009. *Report on Health Related Quality of Life and Lifestyle of Bangladeshi Migrants in Melbourne: Use of MAU instruments, Research Paper 44*. Melbourne: Centre for Health Economics, Monash University.

Khan, M. A. and J. Richardson. 2011. *A comparison of 7 instruments in a small, general population, Research Paper 60*. Melbourne: Centre for Health Economics, Monash University.

Misajon, R., G. Hawthorne, J. Richardson, J. Barton, S. Peacock, A. Iezzi, and J. Keeffe. 2005. “Vision and quality of life: The development of a utility measure.” *Investigative Ophthalmology & Visual Science* 46(11): 4007-15.

Richardson, J. 2010. *Psychometric Validity and Multi Attribute Utility (MAU) Instruments, Research Paper 57*. Melbourne: Centre for Health Economics, Monash University.

Richardson, J., N. A. Day, S. Peacock, and A. Iezzi. 2004. “Measurement of the quality of life for economic evaluation and the Assessment of Quality of Life (AQoL) Mark 2 Instrument.” *Australian Economic Review* 37(1): 62-88.

Richardson, J., G. Elsworth, A. Iezzi, C. Mihalopoulos, I. Schweitzer, and H. Herrman. 2011a. *Increasing the Sensitivity of the AQoL Inventory for Evaluation of Interventions Affecting Mental Health, Research Paper 61*. Melbourne: Centre for Health Economics, Monash University.

Richardson, J., A. Iezzi, S. Peacock, K. Sinha, R. Misajon, and J. Keeffe. 2011b. *Utility weights for the Vision Related Assessment of Quality of Life (AQoL) 7D instrument, Research Paper 67*. Melbourne: Centre for Health Economics.

Richardson, J., J. McKie, and E. Bariola. 2011c. *Review and Critique of Related Multi Attribute Utility Instruments, Research Paper 64, (Forthcoming in A Culyer (ed), Encyclopedia of Health Economics, Elsevier Science San Diego)*. Melbourne: Centre for Health Economics, Monash University.

Richardson, J., J. McKie, and E. Bariola. 2011 forthcoming. “Multi attribute utility instruments and their use.” In *Encyclopedia of Health Economics*, edited by T. Culyer. San Diego: Elsevier Science.

Richardson, J., K. Sinha, A. Iezzi, and M. Khan. 2011d. *Modelling the Utility of Health States with the Assessment of Quality of Life (AQoL) 8D Instrument: Overview and Utility Scoring Algorithm, Research Paper 63*. Melbourne: Centre for Health Economics, Monash University.

Streiner, D. and G. R. Norman. 2003. *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford: Oxford University Press.