MONASH University
Business and Economics

Centre for Health Economics

Research Paper 2011 (64)

Review and Critique of Health Related Multi Attribute Utility Instruments

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

Dr John McKie
Senior Research Fellow, Centre for Health Economics
Monash University

Emily Bariola
Research Assistant, Centre for Health Economics
Monash University

June 2011

Centre for Health Economics
ISSN 1833-1173
ISBN 1 921187 63 8

Correspondence:

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

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

ABSTRACT

Multi attribute utility (MAU) instruments are questionnaires relating to an individual's health and quality of life (HR-QoL). They provide a formula for calculating a utility score from the answers. The utility values may be used in economic evaluation of health related programs.

A small number of MAU instruments dominate the literature. Their history, construction and use are described here. Despite sharing a common purpose, instrument structure, context and scores differ significantly, reflecting different disciplinary traditions and approaches to measurement. This implies that, at present, the outcome of economic evaluations may depend upon the choice of the MAU instrument.

The present paper is a draft entry for the On-line Encyclopedia of Health Economics, edited by AJ Culyer (forthcoming 2014). It provides references omitted from the final text.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Boxes

Review and Critique of Health Related Multi Attribute Utility Instruments

1 Introduction

A multi attribute utility instrument (MAUI) produces a utility score for every health state included in a generic (general) description of the health-related quality of life (HR-QoL). Each of the six instruments reviewed here consists of two parts: a questionnaire and a scoring algorithm which converts responses to the questions into a unique score.

Box 1 illustrates this. The EQ-5D MAU instrument consists of five single 'items', ie questions and response levels (see Box 2). The EQ-5D consists of five single items each relating to a separate dimension of health (the 'descriptive system' or classification). It combines these using the algorithm – formula – shown below the questionnaire. An individual ticking level 1 for each item, ie (1, 1, 1, 1, 1) would obtain a utility score of 1.00; a person scoring (3, 3, 3, 3, 3) the 'all worst' health state would obtain a utility of -0.594. Someone ticking (1, 1, 2, 2, 3) would score 0.225.

As health states change (because of a health program) answers change and the MAU instrument predicts a change in a person's utility.

Multi attribute instruments are useful for the description of cross-sectional and longitudinal health states and for clinical purposes. But the production of a utility score allows the MAU instrument to be used, uniquely, to calculate quality adjusted life years (QALYs) for use in economic evaluation – specifically cost utility analyses (CUA) and, less commonly, for the estimation of QALY based burdens of disease.

Construction of MAUIs: Construction of an MAUI requires three steps. First, the questionnaire ('descriptive system', 'classification' or 'descriptive instrument') is created. Secondly, individuals are interviewed to obtain numerical values relating to the description. Thirdly, a 'model' is used to extrapolate and interpolate the numerical values to all of the possible health states. The third step is necessary because (with one exception) the number of health states described by an MAUI is too large for each to be separately evaluated.

Different MAUI have approached these three steps differently. 'Health states' can be variously described and different MAUI have adopted different theories and definitions of 'health'. Numerical utility scores have been obtained for health states using different 'scaling' techniques and, in particular, with the time trade-off (TTO), standard gamble (SG) and rating scale (RS) which is a visual analogue scale (VAS). Other techniques are available but less commonly used. The models used to extrapolate results can employ statistical techniques, sophisticated averaging or a combination of these.

Box 1 EQ-5D Descriptive system

EQ-5D Descriptive system

1. Mobility (MOB)

MOB 1	No problems walking about
MOB 2	Some problems walking about
MOB 3	Confined to bed

2. Self-Care (CARE)

CARE 1	No problems with self-care
CARE 2	Some problems washing or dressing self
CARE 3	Unable to wash or dress self

3. Usual Activities (ACT)

ACT 1	No problem with performing usual activities (eg work, study, housework, family or leisure activities)
ACT 2	Some problems with performing usual activities
ACT 3	Unable to perform usual activities

4. Pain/Discomfort (PAIN)

PAIN 1	No pain or discomfort
PAIN 2	Moderate pain or discomfort
PAIN 3	Extreme pain or discomfort

5. Anxiety/Depression (DEP)

DEP 1	Not anxious or depressed
DEP 2	Moderately anxious or depressed
DEP 3	Extremely anxious or depressed

Combinations of answers ('Health states') = 3 x 3 x 3 x 3 x 3 = 243

EQ-5D Scoring formula

Utility = 1 - [(.069 MOB2 + .314 MOB3) + (.104 CARE2 + .214 CARE3)
+ (.036 ACT2 + .094 ACT3) + (.123 PAIN2 + .386 PAIN3)
+ (.071 DEP2 + .236 DEP3) + (.081 ANY(A) + .269 ANY(B))]

where [MOB2, ... PAIN3] = 1 (or 0.00) if the respondent did (did not) tick the corresponding response level of the item

ANY(A) = 1 if any level $\neq$ 1; ANY(B) = 1 if any level = 3

Note: The derivation of the formula and parameters (0.69, 0.314 etc) are explained in the text

QALYs and MAUI's: Quality adjusted life years (QALYs) are obtained by multiplying the utility of a health state by the time in the health state (Torrance 1986). For example, if someone spent 10 years in the EQ-5D health state (11223) which is used in the example above, they would gain 10 x 0.225 = 2.25 QALYs.

The use of the EQ-5D or another MAU makes the calculation of utility scores very simple: a questionnaire is distributed to the people of interest and they tick the response category which best describes their health. A problem with this approach – discussed below – is that the MAUI may not allow a very accurate description of a health state. A second approach to constructing QALYs – the 'holistic' or 'composite' approach is to interview the people of interest and construct a brief description of their health state (a 'vignette' or 'scenario'). This is then evaluated by another group using one of the scaling techniques (the TTO, SG or VAS).

The two approaches are different in application, but, in principle, they involve the same steps. A description is obtained from an interview (holistic) or by completing the questionnaire (MAUI). The valuation is carried out in a second interview and with the direct use of a scaling instrument (holistic) or indirectly using the algorithm (MAUI).

Each method has its advantages and disadvantages. MAUI's are cheap and easy to use. Only the questionnaire needs distribution and completion. This facilitates repeated use through time to track changes in utility or to produce a profile of the instrument's dimensions through time.

However, MAUI's have limitations. The health state description is constrained by the content (sensitivity) of the instrument's descriptive system and by the validity of the utility scores produced by the algorithm. The utility index applies to a fixed point in time (subsequently multiplied by life years to obtain QALYs) as distinct from the varying time period which may be embodied in the holistic calculation of utilities (as, for example, with the Healthy Year Equivalent (HYE)). While holistic descriptions are flexible this makes comparison and validation of utilities problematical as there are no agreed norms for the framing and boundaries of the scenarios to be evaluated.

Box 2 MAU instrument related terminology

Algorithm	(or formula) The rule for converting answers to a questionnaire into a number. It is constructed by 'scaling' a 'model'
Attribute	A characteristic or property which an instrument seeks to describe eg vitality, depression, mobility
Construct	An attribute which is constructed or conceptualised as part of a theoretical explanation
Content	The scope and detail of the instrument's descriptive system: the behaviours, outcomes or states which determine an instrument's score
Descriptive system	(or descriptive 'classification; or descriptive 'instrument) The collection of items and dimensions which describe the health state
Dimension	A collection of attributes with a common theme (a 'super construct') eg physical, mental or social health. It usually consists of more than 1 item
Element	a single idea or attribute embodied in an item or dimension eg contentment or exhilaration but not contentment and exhilaration
Instrument	A questionnaire with an associated method for attaching a numerical value to the answers
Item	A linguistic statement generally consisting of a stem (eg 'in the last 7 days I was: …) plus a number of ordered response levels (eg 'always happy' … 'never happy')
Model	A conceptual or mathematical framework which defines how values will be combined (for example, simple or weighted averaging of the level of the item responses)
Reliability	See Box 5
Scaling	(or calibrating) The process of creating the algorithm for attaching numbers to health state descriptions. It requires a scaling instrument (eg TTO or SG) plus a model for combining the numbers produced by the scaling instrument
Sensitivity	The extent to which the instrument content allows the detection of changes in a health state
Validity	See Box 5

Six MAUIs are reviewed in the present paper (Box 3). Their chronology, characteristics and construction are described and compared in Section 2. Section 3 summarizes their use and recognition by health authorities. The different estimates of utility produced by the instruments (Section 4) imply that some, or possibly many, of the utilities currently used are invalid. Part of the reason for the difference may be found in the different theoretical traditions embodied in the instrument. Theory and evaluation are discussed in Section 5. Challenges to the field are outlined in a concluding Section 6.

Box 3 Six multi attribute utility instruments and country of origin

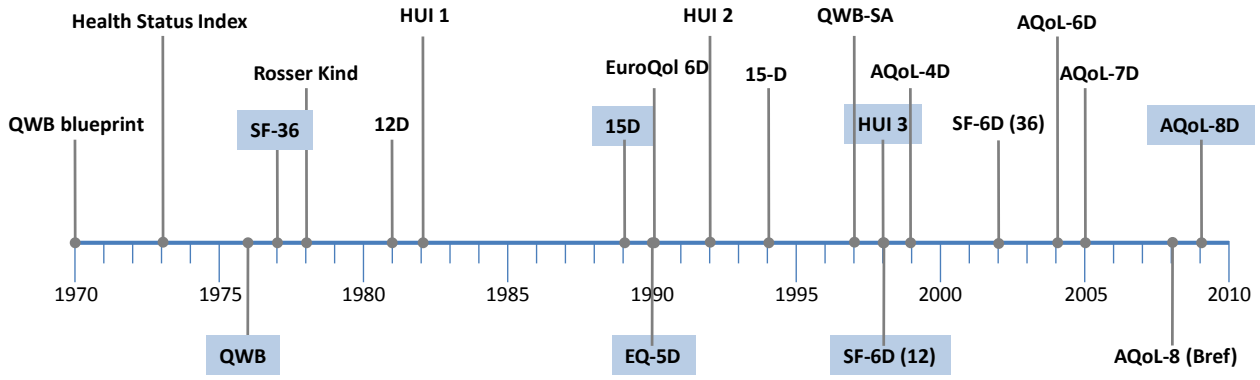
QWB	Quality of Wellbeing Index	... USA
15D	15 dimension instrument	... Finland
EQ-5D	Originally EuroQol (RS and TTO versions)	... Europe/ UK
HUI	Health Utilities Index, 3 versions, HUI 1-3	... Canada
SF-6D	Short form 6D (SF-6D (12) and SF-6D (36)	... UK/USA
AQoL-8D	Assessment of Quality of Life (8D)	... Australia

2 Chronology description and construction of MAUI

2.1 Chronology

Box 3 and Figure 1 document the history of the six MAUIs. Most writers in the area commence with a reference to the famous WHO definition in 1948 of health as a 'state of complete mental and physical wellbeing and not merely as the absence of disease and infirmity' (WHO 1948). This legitimized the concept of 'health' as a single construct. However, it did not provide a basis for measurement. In the USA the 'blueprint' for measurement was published in 1970 by Fanshel and Bush (1970). This provided the theoretical basis for the earliest instruments, the Health Status Index (Patrick, Bush et al. 1973), the QWB (Kaplan, Bush et al. 1976) and the SF-36 (Stewart, Ware et al. 1977). The latter was the empirical basis for the two UK versions of the SF-6D, one directly derived from the SF-36 (Brazier, Roberts et al. 2002) and one from its reduced form, the SF-12 (Brazier, Roberts et al. 2004). The work by Bush and the RAND team was also influential in the construction of the HUI instrument (Kaplan 2005).

Figure 1 History of MAU instruments



In the UK and Europe two separate initiatives resulted in the Rosser Index, initially for hospitals (Rosser and Watts 1972) and subsequently generalised to a 29 health state classification the 'Rosser-Kind index' classifications (1978). Secondly, and displacing this index, the EuroQol was created by a group formed in 1987 (EuroQol Group 1990) which subsequently renamed it the EQ-5D. Conversion of the EuroQol into an MAU instrument for general use followed publication of a scaling algorithm at the University of York (Dolan, Gudex et al. 1995). Earlier Sintonen had created the 12D instrument in Finland and the publication of the revised 15D occurred immediately before the work of the EuroQol Group (Sintonen and Pekurinen 1989).

The three Canadian instruments (HUI) were initiated by Torrance for the evaluation of neonatal intensive care (Boyle, Torrance et al. 1983). This was modified for use in childhood cancer (HUI 2) (Torrance, Feeny et al. 1996) and further developed and scaled for the adult population in the HUI 3 (Feeny, Furlong et al. 2002).

The Australian AQoL instruments were initiated by Richardson. The AQoL-4D was published in 1997 (Hawthorne, Richardson et al.) and subsequently modified as the AQoL-6D in 2004 (Richardson, Day et al. 2004). An additional dimension (VisQoL) was added to measure sensitivity to vision related health states which resulted in AQoL-7D (Misajon, Hawthorne et al. 2005; Peacock, Misajon et al. 2008). Two dimensions were added from a mental health instrument (PsyQoL) to create the AQoL-8D (Richardson and Khan 2009; Elsworth, Richardson et al. 2011).

2.2 Description

Tables 1-3 summarise the MAUIs. Two broad approaches to description ('conceptual type') have been used (Table 1). Following the WHO typology health problems result in impairment, disability and handicap; roughly, body malfunction, limitations of body performance, and problems affecting life in a social context, respectively. Three MAUI have based their descriptions primarily on the last concept (EQ-5D, SF-6D, AQoL). The classification however is imperfect and pain (disability) is also included. Two MAUIs have adopted a 'within-the-skin' approach (disability) – 15D and HUI – although 15D was modified to include one handicap dimension (usual activities). The QWB spans all concepts.

The resulting instruments have between 5 and 15 dimensions with one item per dimension in HUI, 15D, EQ-5D and SF-6D and an average of 4 items per dimension for AQoL-8D. QWB has 3 basic dimensions supplemented with 35 symptom/problem groups which transcend dimensions. Item response levels in the instruments vary from 3 to 6 resulting in between 243 health states (EQ-5D) and 2.37×10^{23} (AQoL-8D). Larger instruments, particularly AQoL, define numerous 'empty' states ('bedridden' and 'no problems with self care').

Dimensions overlap imperfectly (Table 1). Several are unique to a particular instrument and similarly named dimensions include different items. Consequently, to appreciate instrument content requires examination of the items. From Table 2 these vary significantly, in part because of the differing conceptual bases and in part from the level of detail of the instrument descriptions. In principle smaller instruments may indirectly capture the information content of omitted items. Alternatively, they may be omitting content to achieve some other goal (brevity). However, the differences are potentially important for instrument validity and are discussed further below.

Table 1 Instrument descriptive systems

	QWB	15D	EQ-5D	HUI 3	SF-6D	AQoL-8D
Descriptive system						
Conceptual type	Handicap Disability Impairment	Disability (handicap)	Handicap (disability)	Disability	Handicap (disability)	Handicap (disability)
Selection of content	Medical literature matched with Health Interview Surveys	Medical + psychometrics	Consensus	Survey; importance ranking	SF-36, SF-6D, psychometrics	Focus groups, medical and psychometrics
Dimensions	3 + 27 symptoms/problems	15	5	8	6	8
Items		15	5	8	6	35
Response levels	2, 3 (2)	4-5	3	5-6	4-6	4-6
States defined	945	3.1×10^{10}	243	972,000	18,000	2.37×10^{23}
Dimension type						
Physical	27 symptoms +	5 unique +				
Mobility/activity	* *	* *	* *	*	* *	*
Self care			*			
Dexterity				*		
Energy		*			*	*
Cognition				*		
Pain		*	*	*	*	*
Senses (1)		***		***		*
Psycho-social						
Social function	*	*			*	*
Mental function				*		
Mental health		**	*		*	**
Satisfaction						*
Completion time	na	4 minutes	1 minute	3 minutes	2.5 minutes	5.5 minutes
Cronbach's α (2)	0.94[3]	0.81	0.69	0.74 0.81		Dimensions 0.82-0.92 AQoL-8D 0.97
Test-retest (P)	0.93-0.98(3)	0.9 -0.94(4)	0.73(4)	0.77(4)	0.88(4)	0.91-0.89(5)

Notes:
Stars indicate items
(1) Vision, hearing, speech; (2) Ref [x, y, z]; (3) consecutive days; (4) 2 weeks, 4 weeks; (5) 2 months***

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

		Number of symptoms (.) and items (*)					
	Dimension	**QWB[1]**	**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	..	*		*		*
Psycho-social	Sleeping	.	*				*
	Psychological: Depression/Anxiety/ Anger	….	***	*	*	*	*******
	General Satisfaction						****
	Self Esteem						**
	Cognition/Memory Ability	.			*		
	Social Function/ Relationships					*	******
	(Family) Role					*	*
	Intimacy/Sexual Relationships	.	*				*
			15 items	5 items	8 items	12 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.

Table 3 Properties of the combination model and the predicted utilities

	QWB	15D	EQ-5D	HUI 3	SF-6D	AQoL-8D
Theory [(1)]	MAUT	MAUT	Statistical	MAUT	Statistical	MAUT/ statistical
Model type	Additive	Additive	Additive	Multiplicative	Additive	Multiplicative/ exponential
Scaling[(2)]	RS	RS	TTO; RS	SG/RS	SG	TTO
Best health [(3)]	1.00	1.00	1.00	1.00	1.00	1.00
Worst health [(3)]	0.320	0.11	-0.59	-0.36	0.203	-0.04
Utility at Age 1 [(4)]						
34-44	0.67[a]	0.95	0.89 [a]	0.83 [a]	0.80 [a]	0.81 [(b)]
60-64	0.64 [a]	0.87	0.86 [a]	0.80 [a]	0.78 [a]	0.84 [(d)]
Test-retest [(5)] (correlation)	0.59[(1)]	Very high[(3)]	0.61	0.75	0.66[(c)]	0.89[(d)]

Notes:

(1) MAUT = MAU Theory; (2) RS = Rating Scale; TTO = time trade off; SG = Standard Gamble; (3) Best/worst health utilities which are theoretically possible in the model; (4) Values predicted for the general population [(a)] US data n = 462 (35-44); 965 (65-74) (Fryback, Palta et al. 2010) [(b)]Australian data n = 225 (35-44); 340, (60+) (Hawthorne, Richardson et al. 2001) ; (5) (intra-class) correlation between scores obtained after [(c)] 5 months and [(d)]1 month

2.3 Instrument construction

Construction requires three key decisions: (i) how to create the descriptive system, (ii) which scaling instrument and survey methodology to employ, and (iii) which model to use to create an algorithm for extrapolating results.

Quality of Wellbeing Index: The three multi response items of the QWB (mobility, social and physical activity) define 47 health states. In combination with 27 symptom/problem groups this rises to 945 states (Table 2). While these contain no explicitly mental health dimensions the instrument has been used for patients with psychiatric problems.

The QWB descriptive system was derived from the Health States Index (Kaplan, Ganiats et al. 1998). Items were selected using medical references matched against health surveys and particularly the NCHS Health Interview Survey. The descriptive system was based upon 343 'core descriptions' (items) and scaled using VAS responses from the general population of San Diego (n = 866). An additive algorithm was used of the form:

$$VALUE = 1-D_1-D_2-D_3-S$$

where D_i are the dimension scores and S is the score for the worst symptom. Distribution of scores for the general population are approximately normal. Perfect scores are rare and there are neither significant ceiling nor floor effects.

QWB was the first MAUI. Originally administered by trained interviewers, a self-administered version (QWB^{SA}) was created in 1997 (Andresen, Rothenberg et al. 1998). Translations exist into Spanish, German, Italian, Swedish, French-Canadian and Dutch. Information and the user annual may be obtained at https://hoap.ucsd.edu/qwb-info/

15D: The descriptive system of the 15D has 15 items, 14 relating to disability (mobility, mental function, etc) and one to handicap ('usual activities'). The instrument was based upon a review of the Finnish health policy documents. The resulting 1981 version was subsequently revised following feedback from the medical profession in 1986 and further revised in 1992 following user feedback and factor analysis

(Sintonen 1994a). An additive model with VAS scaling was used. Five separate weighting systems were compared by using responses from five Finnish population samples (n = 2500) and transformations of VAS data into 56 'utilities' using an econometric transformation. Results demonstrated convergent validity of 15D values (Sintonen 1994b).

The 15D has been modified for children (16D) and has been translated into 25 languages with 4 in preparation. The 15D website is: http://www.15d-instrument.net/15d

Health Utilities Index (HUI): HUI 3 consists of 8 items with either 5 or 6 levels. The descriptive system is a modification of HUI 2 and reflects the importance ranking assigned to a list of 15 symptoms in a Canadian survey of parents by Cadman and Goldsmith (Feeny 2002). The 'within-the-skin' – ie disability based – descriptive system has no social or handicap based dimensions (Torrance, Boyle et al. 1982). VAS scaling was used with 504 adults from Ontario, Canada and the scores were converted to a standard gamble (utility) using the power function fitted to 3 points. The HUI combination model was based upon the assumption of structural independence and employs the multiplicative model recommended by Decision Analytic MA (Multi Attribute) theory (Feeny 2002). Empirically the correlation between items varies between 0.02 and 0.35 which is consistent with the conventional psychometric definition of independence.

HUI questionnaires are available in English, Chinese, Japanese, Russian, Dutch, French, German, Italian, Portuguese, Spanish, Czech, Polish, Finnish, Norwegian and Danish. Sixteen versions of English are based on mode of administration, assessment viewpoint and duration of assessment period. The website is http://fhs.mcmaster.ca/hug/.

EQ-5D: The 5 item 3 level EQ-5D defines 243 health states. It was originally designed to compare broad preference patterns across Europe and not as a stand-alone MAUI for economic evaluation (Sintonen, Weijnen et al. 2003). The original EuroQol Group considered it 'highly unlikely that such a simple instrument could be comprehensive' (Brooks and EuroQol Group 1996). Following the development of preference weights at the University of York (Dolan 1997) it became widely accepted as a generic MAUI and eventually became the preferred instrument by the UK National Institute of Health and Clinical Excellence (NICE). The UK weights, which are the most widely used, employed VAS and TTO data from a survey of 2997 members of the UK population. The main results of the econometric analysis are reported in Box 1. Models were also created for different socio demographic groups with 8 algorithms estimated using both TTO and VAS. The TTO algorithm for the general population is most commonly used.

The correlation between EQ-5D dimensions varies, typically from about 0.24 to 0.64 (Feeny 2002) indicating structural dependence. However econometric scaling was used to combine items which eliminates 'double counting' at the mean of the sample.

The EQ-5D has been translated into 150 languages. A version for children aged 7-12 years has been translated into 12 languages. An algorithm has been estimated in the USA using data from 3773 respondents (Shaw, Johnson et al. 2005). In 2009 the EQ-5L, a 5 response level instrument (with the same items) was published and the Group Executive approved the use of 'bolt-ons' to increase instrument sensitivity for particular health states. The website is http://www.euroqol.org/.

SF-6D: Two versions of the SF-6D instrument were derived; one from the SF-36, the most widely used generic HRQoL instrument, and the other from its derivative, the SF-12. Consequently, utility scores may be derived from any study reporting values from these instruments. 'SF-6D (12)' and 'SF-6D (36)' are similar except for a reduction in the response categories for two items in SF-6D (12) which reduces the possible health states from 18,000 to 7,500.

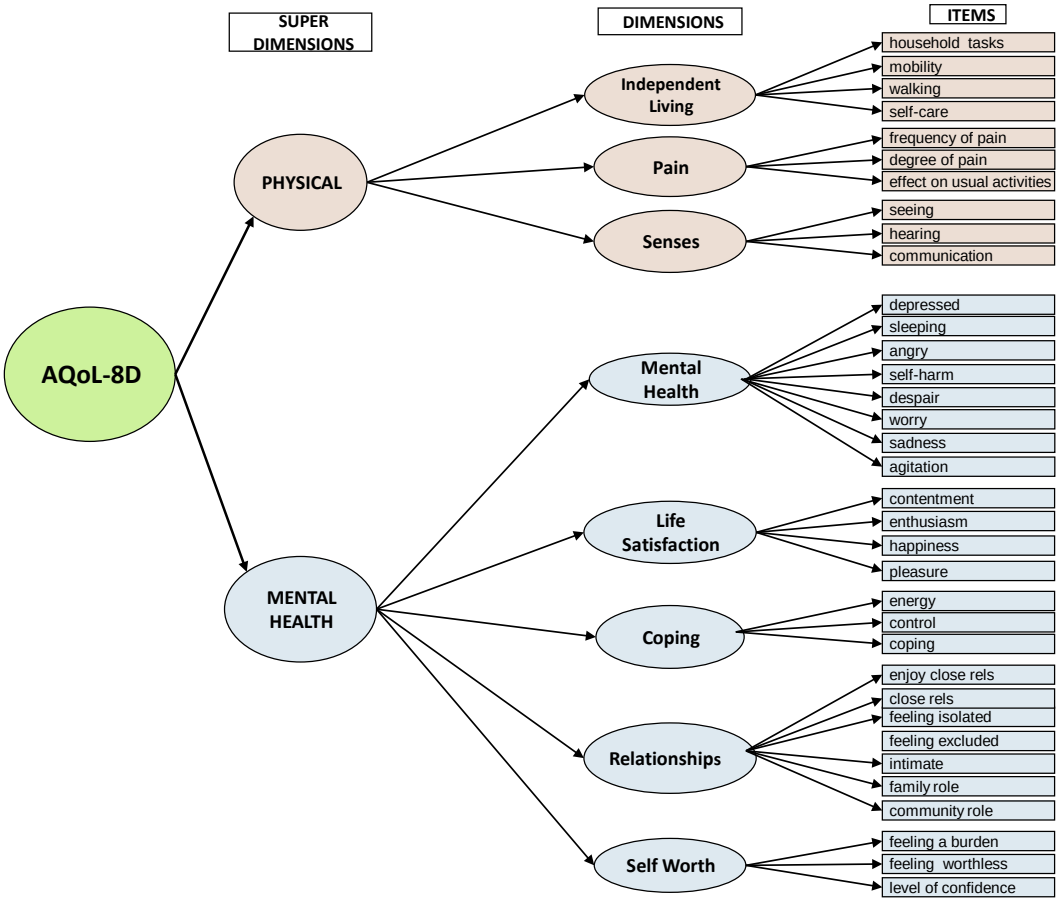
The items of the descriptive system were derived from the factor analysis and psychometric properties undertaken in developing of the SF-36.

Utility scores were obtained using the standard gamble to evaluations of 249 health states with 6 observations from each of 611 UK participants. Initial econometric modelling used random effects linear regressions on mean health state values. Re-estimation using rank data subsequently gave similar results. Non-parametric Bayesian approach achieved greater predictive power and reduced the minimum predicted value from 0.301 to 0.203. This algorithm is now recommended.

Versions of the instrument have been developed in Australia, Brazil, Hong Kong, Japan, Portugal and Singapore. The website is: http://www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d.

Assessment of Quality of Life (AQoL): The AQoL descriptive systems were constructed from reviews of instruments measuring theoretically indicated health dimensions, from focus groups and from 'construction surveys'. These administered large numbers of items to selected patients and the public. Multiple items were selected per dimension using factor analyses and SEM (Structural Equation Modelling). A multi-level model was adopted which first combines items into dimensions and secondly combines dimensions into the overall AQoL model. The structure of AQoL-8D is shown in Figure 2.

Figure 2 Structure of the AQoL-8D



To overcome the effects of structural dependence between items, AQoL-4D sought orthogonality according to psychometric norms between dimensions and combined items and dimensions using a multiplicative (KDA) formula. Subsequent AQoL's dropped the attempt to achieve orthogonality as it proved too restrictive. Rather a stage 2 econometric correction was introduced in which the TTO values of holistic states were regressed upon the stage 1 multiplicative scores for dimensions. Exponential models were employed. AQoL-8D introduced a similar 'correction' in the estimation of each dimension using independent valuations of holistic dimension scores.

AQoL-8D used a sample of 712 to construct the descriptive system and a second population sample of 628 to obtain TTO scale values (322 patients, 306 other). The scaling survey obtained values for 162 multi- item dimension health states and 375 multi-dimensional health states from 629 respondents, half patients and half from the general population.

Transformations have been created between AQoL-4D, 6D and 8D. AQoL-4D (the original 5D instrument scaled without the original dimension for symptoms) has been reduced to an 8 item AQoL-Bref or AQoL-8 (which should not to be confused with the AQoL-8D). The four AQoL instruments have been translated into traditional and simplified Chinese, Spanish, German, Arabic, Norwegian and Danish. The AQoL website is: http://www.aqol.com.au/.

3 Instrument use and acceptance

3.1 Instrument Use

Information on the use of each of the MAU instruments was obtained from the Web of Science database for the period 2005-2010 (Web of Science 2011) and supplemented by references provided to the authors or from the instrument websites. These were used to construct Tables 4-7. The search identified 1682 studies which employed one of the MAU instruments.

Table 4 indicates that EQ-5D was the most popular instrument by a significant margin, with 63.2 percent of the 1682 studies using it. This was followed by HUI 3 (9.8 percent) and SF-6D (8.8 percent). At the other end of the scale 15D and AQoL were included in 6.9 and 4.3 percent of studies respectively and the QWB, the earliest widely used instrument, accounted for only 2.4 percent of total use.

The EQ-5D also dominated use in most countries and was only exceeded in Canada by the HUI 3 and in Finland by the 15D. Table 4 reveals significant 'local loyalty' with the use of all instruments peaking in their country of origin. Apart from EQ-5D, only HUI 3 and SF-6D achieved significant use in other countries.

Use of the instruments was also very concentrated. European countries accounted for 55 percent of usage and the addition of usage in the USA and Canada raises this to 80.5 percent. Within Europe use was also concentrated, with Finland and Netherlands each accounting for more than 8 percent of the total, or double the usage by Germany, despite its much larger population, and over 65 percent of the usage by all other European countries combined. The extent to which this is attributable to language and publication bias is unknown.

Only 15 percent of the studies included in Table 4 were primarily concerned with economic evaluation, (which need utility scores) as distinct from their use as generic tools for the measurement of HRQoL (which do not require scores to be 'utilities'). The disease categories in which they were used are reported in Table 5. This reflects a broad acceptance of MAU instruments across the spectrum of

disease categories, possibly reflecting the widespread use of self-reported disease specific instruments in medicine. Given the scope of the literature search, however, the number of studies published in most of the disease areas is relatively small.

Table 4 Number of studies using the 6 MAU instruments

Instrument	Country of Study Population												Economic evaluation	Total studies	%
	USA	Canada	UK	Finland	Germany	Spain	Sweden	Netherland	Other Europe	-Australasia	Multi-nationals	Other			
QWB	31	4			1		1					4	6	41	2.4
15D		1		93	1		3		17	1			18	116	6.9
EQ-5D	133	52	181	24	62	57	67	103	181	34	97	72	166	1063	63.2
HUI 2	27	25	9		3		1	7				6	8	78	4.6
HUI 3	43	60		21	3		2	15	2	2	6	10	22	164	9.8
SF-6D	30	16	27	1	2	6	2	16	23	6	6	13	27	148	8.8
AQoL*		1								69	1	1	6	72	4.3
Total	264	159	217	139	72	63	76	141	223	112	110	106	253	1682	100
%	15.7	9.5	12.9	8.3	4.3	3.7	4.5	8.4	13.3	6.7	6.5	6.3	15.0	100.0	

Table 5 MAU Instrument use by disease sub-group 2005-2010

Disease Sub Groups	QWB	15-D	EQ-5D	HUI 2	HUI 3	SF-6D	AQoL	Total	%
Muscular skeletal	4	12	107	3	4	17	5	152	9.1
General population	7	4	87	18	19	15	4	154	9.3
Cardio	2	15	84	4	7	10	11	133	8.0
Arthritis	0	8	71	5	11	21	5	121	7.3
Cancer I	4	6	69	5	16	2	2	104	6.3
Degenerative and Elderly	3	3	69	6	14	3	8	106	6.4
Internal organs	0	10	67	5	5	10	1	98	5.9
Psychiatric	3	8	66	2	6	8	6	99	6.0
Diabetes mellitus	1	2	51	2	10	2	1	69	4.1
Other	1	2	51	3	7	8	1	73	4.4
Medical patients	3	8	49	3	11	9	1	84	5.1
Injury	1	6	44	3	4	6	8	72	4.3
Eating/ Obesity	2	5	29	0	2	6	5	49	2.9
Respiratory	2	1	27	2	6	2	0	40	2.4
Vision	1	4	26	0	6	1	0	38	2.3
Neurological	0	8	20	6	7	0	3	44	2.6
Skin	0	0	20	0	1	1	0	22	1.3
Female conditions	2	4	19	1	4	3	1	34	2.0
Trauma	0	0	19	0	0	2	3	24	1.4
Chronic condition	0	3	17	0	4	2	0	26	1.6
HIV	1	1	15	1	4	2		24	1.4
ENT	3	2	15	6	11	2	0	39	2.3
Renal	1	3	11	1	2	6	1	25	1.5
Autoimmune	0	1	9	0	2	7	3	22	1.3
Rheumatic		1	5	1	1	2		10	0.6
Total	41	117	1047	77	164	147	70	1663	100.0
%	2.5	7.0	63.0	4.6	9.9	8.8	4.2	100.0	

Notes: AQoL studies include 61 AQoL-4D, 7 AQoL-6D and 2 AQoL-8D

3.2 Acceptance by Health Authorities

The different instruments enjoy varying degrees of acceptability by health authorities and several are mentioned in national pharmaceutical guidelines or draft guidelines (Box 4). In the UK, The National Institute for Clinical Excellence (NICE) recommended use of the EQ-5D, while acknowledging that it 'may not be an appropriate measure of health-related utility in all circumstances'. It has been used to establish levels of population health in Spain (1994 Catalan health survey interview), the UK (UK Department of Health Omnibus Sample Survey 1996, Health Survey for England), and the US (Medical Expenditure Panel Survey by the Agency for Healthcare Research and Quality). It has also been used in the NHS PROMs (Patient Reported Health Outcomes) programme, the purpose of which is to enable the patient perspective to inform decision-making within the NHS. The HUI 3 has been included in all of the major Canadian general population health surveys since 1990 (for references see: http://fhs.mcmaster.ca/hug/). The 15D has been included in the Finnish National Health Survey 1995/96, the Health 2000 survey in Finland, and the Danish National Health Survey 2000. AQoL-4D and 6D have been used in the South Australian Health Surveys and Australian National Heart survey.

Box 4 International Pharmacoeconomic Guidelines

References	EQ-5D	HUI	SF-6D	QWB	15D	AQoL
Hungary (Szende, Mogyorósy et al. 2002)	Noted as internationally recommended	Noted as internationally recommended		Noted as internationally recommended		
Poland (Orlewska and Mierzejewski ; Orlewska and Mierzejewski 2003)	Recommended for measuring generic quality of life and the utility of health states.	Recommended for determining the utility of health states.				
(Belgium 2008)	'As long as Belgian valuation sets for other instruments are not available, the use of the Flemish valuations for the EQ-5D health states is recommended'.					
(France 2004)	Recommends QWB, HUI and EuroQoL: 'validations of French versions of the latter two are proposed'.	Recommended		Recommended		
(Netherlands 2006)	Recommended	Recommended				
UK (NICE 2008)	Preferred, but ' may not be an appropriate measure of health-related utility in all circumstances.'					
(Ireland 2010)	Recommended		Recommended			
(Scotland 2007)	Recommended, but 'it would be inappropriate to require the use of the EQ-5D to the exclusion of any other valid generic utility measures.'					
(Sweden 2003)	Recommended as an indirect measure for QALY-weightings.					
Italy (Capri, Ceci et al. 2001)			Recommended			
(Canada 2006)	Noted as widely used	Noted as widely used	Noted as widely used		Noted as widely used	
(USA 2009)	Recommended	Recommended				
New Zealand (PHARMAC 2007)	'The New Zealand EQ-5D Tariff 2 recommended. 'Other instruments can be used, however their use should be well justified'.					
Australia (PBAC 2008)	Acceptable	Acceptable	Acceptable			Acceptable

4 Comparison of instruments

Tables 6 and 7 report the number of comparisons between scales and methods used from 2005 to 2010. Most notable is the predominance of simple correlation studies. Intra-class correlation, the preferred statistic even in simple comparisons, is seldom used and psychometric analyses are rare. This is discussed further below.

There have been surprisingly few multi-MAUI comparisons. In an early Australian comparison 956 hospital and general respondents were administered the EQ-5D, SF-6D, 15D, HUI 3 and AQoL-4D. The proportion of instrument variation explained by other instruments varied from 41-59 percent leaving an average of 44 percent unexplained. The highest explanatory power was achieved by 15D followed by AQoL (Table 8). In a recent US study of 3844 adults, were surveyed to compare the EQ-5D, QWB^{SA}, HUI 2, HUI 3 and SF-6D. A weaker association was found than in Australia (reflecting the use of only general population respondents). Overall 53 percent of instrument variance was not explained (Table 7). Recent work indicates that the strength of the comparison between two instruments is likely to vary across the distribution of health (Seymour, McNamee et al. 2010).

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’.

Using the same data, Fryback et al. (2010) conclude that ‘linear functions may serve as crosswalks (transformations) amongst these indices only for lower health states, *albeit* with low precision ... indices are imprecisely related’ (2010 p5). Using data for 376 cataract patients, Kaplan et al. (2010) examined base line and one month follow-up data using the same five instruments. Their conclusion was that the various MAUI are not equally responsive to change.

Similar lack of concordance has been found in other multi-instrument studies. A comparison of the 15D, EQ-5D and SF-6D in the context of AIDS concludes that different measures give different utility values (Stavem, Frøland et al. 2005). In the context of spine patients lower correlations were found between EQ-5D, SF-6D HUI and QWB than in the USA and the authors conclude that differences in instrument outcomes warrant caution ((McDonough, Grove et al. 2005). The same instruments were administered to a sample of 264 German rehabilitation patients with mild to moderate muscular skeletal cardiovascular and mental health problems. The authors conclude that the instrument values are not equivalent (and) may have considerable effects upon health economic evaluation studies (Mook and Kohlmann 2008). Results of an analysis of 1011 Italian patients who attended GP clinics concluded that agreement between EQ-5D, HUI 3 and SF-6D was ‘quite low’ (Quercioli, Messina et al. 2009 p 390).

Reasons for these differences are discussed in Section 5. However, one proximate cause is the difference in upper end sensitivity as indicated by ceiling effects. Significant differences were found in the early Australian study. 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 3 illustrates instrument validity with results form a more recent Australian study (Khan and Richardson 2009). The data reflects 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 (AQoL and HUI had minimum values of 0.42 and -0.04 respectively). Additionally, at all levels of one instrument there was significant variation in the value of other instruments as with cluster 3 and 4 in the previous figure. When SF-6D = 0.6,

HUI 3 and AQoL-8D values varied from (0.25-1.00) and (0.55-0.95) respectively; when AQoL-8D = 0.8 HUI 3 and SF-6D varied from (0.25-1.00) and (0.10-1.00) respectively. Importantly, differing results were obtained from the same individuals and the magnitude of the problem to be explained is indicated by the extreme range of individual differences and not by average differences in group scores. Some of this variation is random. A small amount can be attributed to the choice of preference instrument; an unknown but large amount must be attributed to the instrument descriptive system and scoring models.

Table 6 Validation Studies (2005-2010) Comparison with other scales

Instrument	Type of Scale [1]			Head to head comparisons [2]							Total MAU comparisons
	Disease specific instrument	Non-utility instrument	Generic MAU instrument	QWB	EQ-5D	SF-6D	HUI 2	HUI 3	15D	AQoL	
QWB	10	0	28	-	7	6	6	8	1	0	28
EQ-5D	137	53	76	7		57	16	26	9	5	120
SF-6D	21	9	57	6	57	-	10	16	3	3	95
HUI 2	22	3	52	6	16	10	-	18	1	0	51
HUI 3	37	11	71	8	26	16	18	-	1	2	71
15D	6	3	15	1	9	3	1	1	-	1	16
AQoL	5	5	11	0	5	3	0	2	1	-	11*
Total	238	84	310	28	120	95	51	71	16	11	392

Notes

(1) Number of separate publications classified by the instrument which was the principal focus of the study

(2) Number of comparisons. Studies with (3+ instruments) are entered multiple (2+) times

(3) Combines AQoL 4D, 8D; 5 studies were pre 2005

Table 7 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	0.58	0.55		0.64
EQ5D		1.00			0.53
HUI 3		0.41	1.00		0.55
SF6D	0.59	0.56	0.44	1.00	0.55
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).

Table 8 Ratio of dimension scores: Individuals above to below predicted utilities on 4 instruments

		Physical dimensions			Mental, Social dimensions					Overall	
		Ind Living	**Pain**	**Senses**	**Mental Health**	**Life Satis**	**Coping**	**Relations**	**Self worth**	**Physical**	**Mental**
Average ratio from 12 regressions		1.05	1.08	1.07	1.13	1.06	1.07	1.08	1.06	1.07	1.07
		Deviation from average ratio									
EQ-5D	HUI	0.0	0.06	0.0	0.08	0.02	0.02	0.05	0.02	0.02	0.04
Predicted	SF-6D	-0.01	0.03	0.08	0.09	0.02	0.15	0.06	0.03	0.02	0.05
By	AQoL-8D	-0.03	0.02	0.03	-0.06	-0.02	-0.05	-0.03	-0.02	-0.02	-0.03
HUI	EQ-5D	-0.01	0.00	0.02	-0.04	0.06	-0.01	-0.05	-0.01	-0.02	-0.02
Predicted	SF-6D	-0.05	-0.08	-0.03	-0.17	-0.07	-0.08	-0.10	-0.05	-0.08	-0.08
By	AQoL-8D	-0.05	-0.03	-0.03	-0.21	-0.07	0.07	-0.16	-0.05	-0.04	-0.10
SF-6D	EQ-5D	0.02	0.00	-0.02	-0.06	-0.01	-0.02	-0.05	-0.01	0.00	-0.02
Predicted	HUI	0.01	0.00	-0.04	-0.07	-0.01	-0.02	-0.01	-0.03	-0.01	-0.02
By	AQoL	-0.03	-0.02	-0.04	-0.12	-0.07	-0.03	-0.08	-0.03	-0.03	-0.05
AQoL	EQ-5D	0.05	0.01	-0.03	0.14	0.07	0.06	0.08	0.04	0.03	0.08
Predicted	HUI 3	0.03	0.01	-0.02	0.18	0.06	0.06	0.15	0.06	0.02	0.10
By	SF-6D	0.03	0.01	0.03	0.15	0.07	0.06	0.10	0.07	0.02	0.11

Table 9 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

Notes

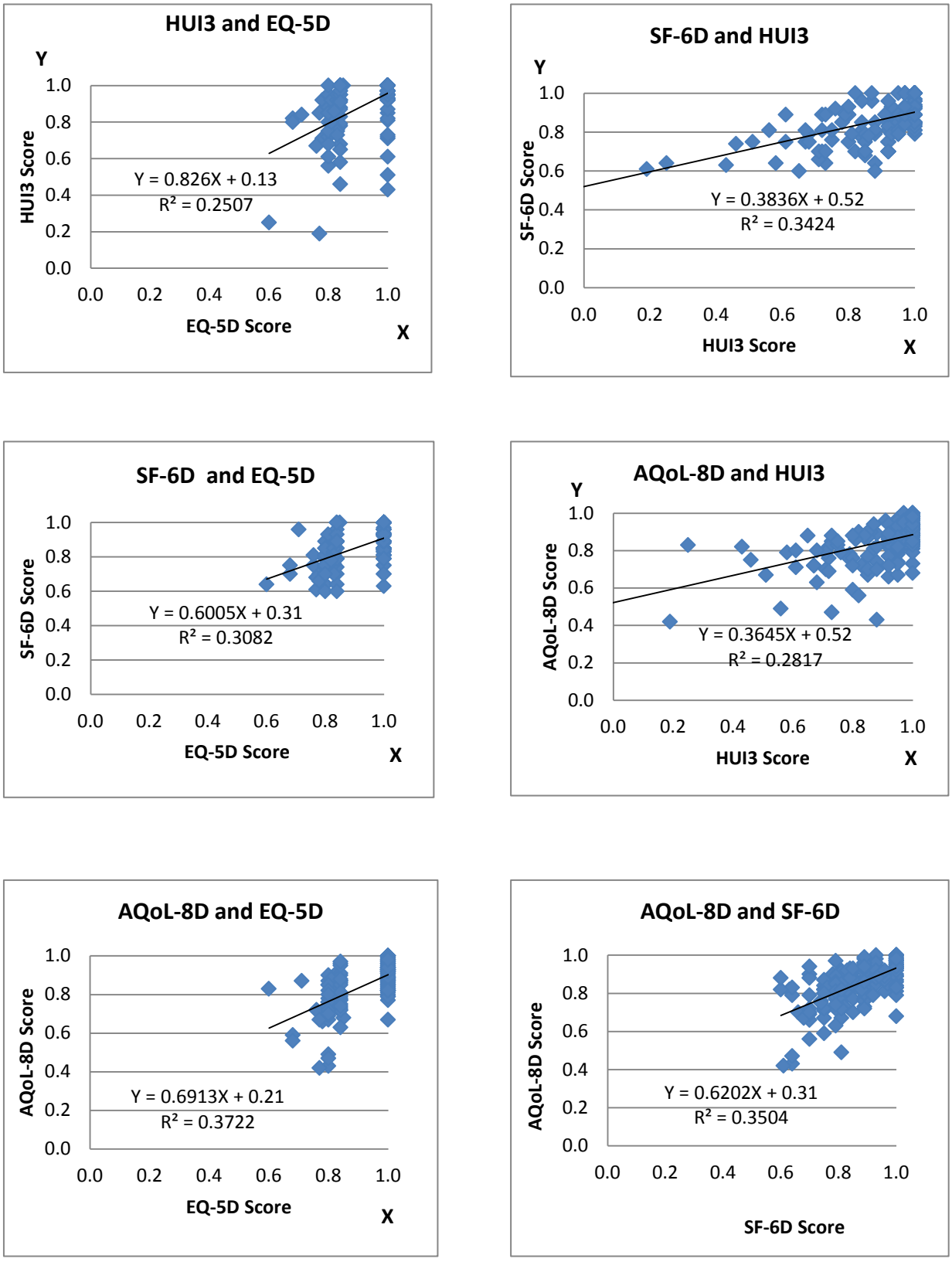
(1) Increase in utility if an individual is cured from the permanent problem and returned to normal or best health

(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 3 Pair-wise comparison of 4 MAU instruments



Source: Khan and Richardson (2011)

5 Theory and evaluation

5.1 Theoretical foundations of MAUI

Current MAUIs draw upon theory from three relatively distinct disciplines: decision analysis, psychometrics and economics/econometrics. The traditions in these are not always consistent, reflecting the problem context from which they arose.

Decision analysis: The 15D, HUI and AQoL all seek theoretical justification, at least in part, from MAU theory, a sub-set of DA theory. This recommends that decisions be analysed in several stages: (i) enumeration of all possible consequences; (ii) uncertainty analysis and construction of a decision tree; (iii) assignment of utility scores to possible consequences; and (iv) maximisation of expected utility. Where the number of choices and outcome are large (such as medical decision making) MAU theory is recommended in the third stage. Consequences should be broken into attributes (dimensions) capable of describing all outcomes, utility scores assigned to each of the attributes and, depending upon the subsequent assumption, an MA combination function employed.

MAU theory requires that descriptive dimensions are structurally independent. A business model optimising output as a function of total revenue, total cost and profit would result in 'double counting' as the last attribute is the sum of the other two. Depending upon the nature of preferences (for the attributes) DA models may be additive, multiplicative or multi-linear, the latter being generally too complex to operationalize. Choice between additive and multiplicative models depends upon the magnitude of the dimension weights. If these sum to unity additive models may be used. If they exceed unity then more complex multiplicative models are applied.

The 15D assumed additive independence. The analyses for HUI and AQoL instruments found dimension preference weights implying multiplicative models (HUI 3 experimented with, but dropped, a partial multi-linear model).

Psychometric theory: Psychometrics is the basis of measurement theory in education and psychology, subjects which, like HRQoL, are concerned with unobserved constructs. Its potential contribution is three-fold: first, it prescribes methods for constructing instruments; secondly, it describes criteria for their evaluation; and, thirdly, it describes numerous forms of bias and other sources of measurement error. In the present context its main message might be the dictum that 'what you measure may not be what you think you are measuring'.

A tension exists between the psychometric and DA approaches. In the former it is assumed that most items being measured correlate to some extent and that the scale for a satisfactory construct requires a minimum of 3 and preferably 4 items for content validity. For example, arithmetic competence might require demonstrated skill in addition, subtraction, multiplication and division. As noted, MAU theory requires orthogonality to avoid double counting of utilities. This problem has received little attention from health economists.

The MAU literature has generally been selective in its use of psychometrics. It has focused on convergent and discriminant validity in testing instruments. With the exception of AQoL, none of the instruments employed psychometric methods to construct the descriptive system. One explanation offered for this is that it is preferences that are important, not description. However valid preference measurement requires valid description. (If a descriptive element is unimportant the preference weight will be zero.) The use of correct preference methods cannot compensate for the absence of a non-trivial descriptive element. Descriptive validity is discussed below.

Economics and econometrics: MAUIs were developed to assist with economic evaluation and specifically the measurement of QALYs. Consequently, the gold standard for evaluating an MAUI is whether or not it measures utility. This implies that a preferences-based instrument should be used for scaling and this is generally interpreted as implying the use of the SG or TTO. However, the subject is controversial, and some argue that there are insufficient reasons for excluding the VAS. Recently, weights have been assigned using ranking techniques (McCabe, Brazier et al. 2006) and Item Response Theory, although the latter technique requires assumptions usually violated in the health sector.

The DA requirement of item orthogonality is difficult to achieve and the resolution of this problem in the EQ-5D, SF-6D and AQoL-8D has been to employ a variety of regression techniques to apportion the contribution of explanatory items to the value of the dependent preference measure. This must be an independently scaled MA health state. The choice of regression model is contentious. From MA theory (above) linear models may be inappropriate if items lack additive independence, but these models are employed in the EQ-5D and SF-6D.

Competing claims have been made about the use of DA and econometric techniques but the evidence is limited. Both approaches are based upon a set of assumptions and constraints which are violated to a greater or lesser extent depending upon the context. This suggests that validation requires context specific evidence.

6 Evaluative criteria

Evaluation criteria proposed in the literature are generally uncontentious. Instruments should be practical and not impose a significant response burden. They should be reliable. (Measurement error should be a small fraction of total variability as judged, for example by a test-retest and Cronbach's alpha). As noted, the MAU instruments reviewed here have evidence of these properties.

The longest instrument – AQoL-8D – takes an average of 5.5 minutes to complete. Test-retest and Cronbach alpha coefficients are satisfactory.

The most contentious criterion is validity: whether or not an instrument measures what it purports to measure. It is contentious because the lack of agreement between instruments noted earlier implies that some or all of the MAUI are not universally valid.

Validity: The concept of validity and its application has been widely discussed in psychometrics but less commonly in health economics. Different types of 'validity' have been variously classified (see Box 5). The common element is that each is a test of the instrument which justifies greater or lesser confidence in its use, depending upon the stringency and outcome of the test. This means that in practice an instrument is never (fully) validated in the sense that it has been 'proven universally correct'. Rather, instruments are more or less supported both empirically and theoretically (an interplay sometimes described as a nomological net). The persuasiveness of the evidence may vary by health state. Importantly, validity is not value free. All of the MAUIs assume that the gold standard for scaling and economic evaluation is individual preferences. The assumptions are not necessary or universally accepted.

Box 5 Validity reliability related definitions

Validity: Measurement of what is intended

Validation: A process of determining (the appropriate level of) confidence in the inferences drawn from instrument values

Construct: A concept created to explain observed relationships

Construct validity: The construct measures what is intended

a) **convergent validity**: correlation with other measures expected to correlate with the construct
b) **discriminant validity**: non correlation with measures of different constructs (eg MAU instruments, blood pressure
c) **discriminative (extreme group) validity**: discrimination between different groups (patients, public)

Content validity: There is a representative sample of target elements in the descriptive system (ie outcomes, behaviours, symptoms, etc) or elements which vary directly with the elements of interest

Face validity: The content appears adequate upon inspection

Criterion validity: Constructs behaviours expected as judged by external criteria

a) **Gold Standard validity**: the instrument correlates with the gold standard measure
b) **Concurrent validity**: the instrument correlates with the criterion
c) **Predictive validity**: the instrument predicts other (criterion) variables as expected

Reliability: A measure of consistency. It is the proportion of the total variability in scores which is accounted for by the differences in the average values across observations. It applies to the interval consistency of the items of an instrument and to the test re-test consistency of the instrument over time.

MAUI Validity: MAUI validity depends upon the validity of its three components: the descriptive system, the scaling method and the combination model. There is no agreement concerning the scaling instrument (TTO, SG, VAS, PTO). However the correlation between instruments is high and could not explain most of the variation between MAUIs.

Combination models differ and the evidence for the assumptions behind them is incomplete. Validity could be tested by comparing the model estimates of health state utilities with independent holistic estimates of the same health state description (ie using the MAUI descriptors). Few and limited studies have been reported. A related test is to ask respondents to think about their own health and to carry out a direct valuation of it using a scaling instrument and to compare the result with the values predicted from an MAUI instrument. Two tests of this type obtained good average but poor individual correspondence between values for the HUI.

Descriptive systems: Descriptive systems differ very significantly in size, item content and syntax. The effect of this on MAUI scores and validity is an unresolved though critical issue. Evidence of descriptive validity may be obtained from three types of validation test, viz (i) content validity (the simplest form of which is face validity) is the requirement that an instrument contains a full sample of a construct's behaviour or states. For example, omitting arithmetic would cast doubt on the content validity of an instrument measuring mathematical ability. Content validity subsumes 'sensitivity' which is an instrument's ability to detect changes in content; (ii) construct validity (the validity of the construct) is tested by discriminant and convergent validity – by an instrument's (lack of) correlation with other instruments which do (not) measure the construct;

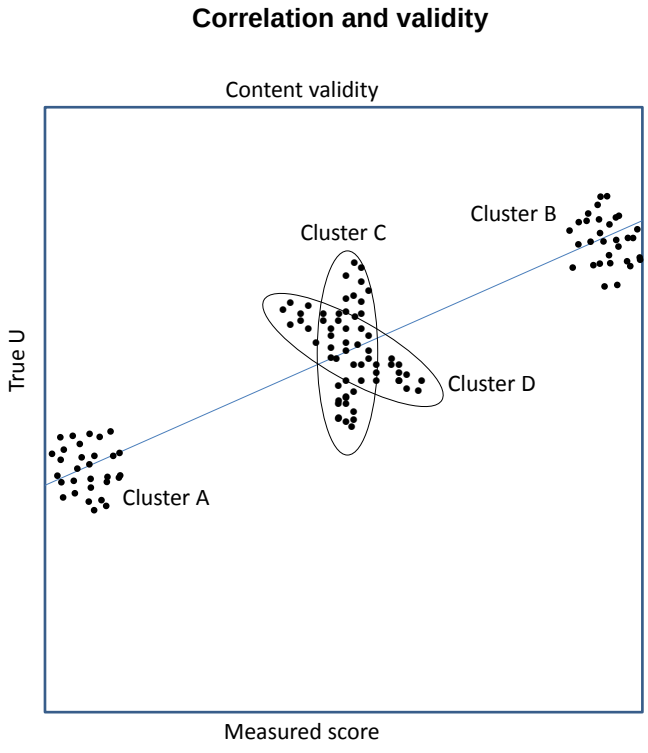
(iii) criterion validity which is referred to below as 'economic or predictive validity' is the ability to predict what is expected when this is independently measured.

Construct validity: The great majority of the studies summarised above in Tables 5 and 6 are concerned with construct and principally convergent validity. This is a weak form of validation for an MAUI. It is necessary but far from sufficient for strong confidence in an instrument. This is particularly true for comparisons with disease specific instruments where simple correlation is the only form of comparison.

Correlation will occur as long as an instrument can, minimally detect extreme values. For example, visual inspection will distinguish obviously very sick and old from young and athletic individuals and, since correlation is disproportionately affected by extreme values, casual inspection will correlate with a gold standard test of health irrespective of insensitivity over most of the scale. In Figure 4 an instrument, I, will correlate with true utility because of clusters A and B irrespective of cluster C and D where the relationships differ from instrument prediction possibly as a result of unmeasured attributes.

A further problem which is illustrated in Figure 4 is that correlation does not indicate that values are similar or that a subset of correlating data will necessarily correlate. In the linear relationship U = a + b I where I is an instrument's estimate of true utility U, instrument validity would imply that a = 0; b = 1.0. For this reason a better measure of association than correlation is the intra-class correlation (ICC) which tests the equivalence of absolute values. In Table 6, however, only a minority of the studies use this technique. The difference is potentially important. In the early Australian five instrument study, the 15D had the highest average correlation with other instruments (construct validity). However incremental changes in 15D were about half the magnitude of corresponding changes in other instruments, indicating a poor ICC.

Figure 4 Insensitivity/content invalidity



Source: Richardson (2010)

Content validity: In contrast with construct validity little has been written in economics about content (or descriptive) validity. From Tables 1 and 2 sources of potential differences are obvious. HUI and QWB have no items relating to self esteem, social or family relations. HUI uniquely contains cognition and dexterity. However it has no dimensions for handicap and mental health respectively. Agreement by the EQ-5D executive to increase the number of response levels and permit 'bolt-ons' represents an attempt to increase the instrument's content validity.

The early five instrument Australian study anecdotally illustrated the importance of descriptive content validity with a respondent's score of 0.14 and 0.8 for the HUI 3 and EQ-5D respectively. When the HUI items for sense perception were altered from their reported to the highest HUI item score (effectively removing senses as a direct source of disutility), the predicted HUI utility score rose to 0.74; that is, 91 percent of the original difference was attributable to items in HUI which are not included in the EQ-5D. This indicates that while general items might in principle, fully capture the content of specific items, the EQ-5D does not do so in the context of sense perception.

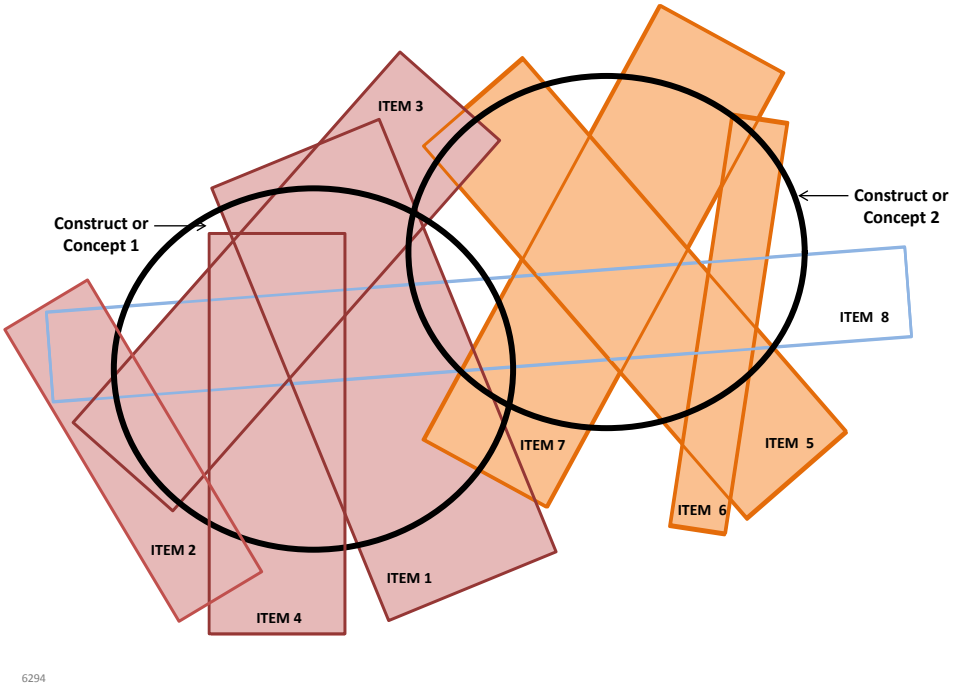
In terms of Figure 4, sense perception might account for cluster C (and other omitted content for cluster D) despite the overall correlation attributable to cluster A and B. The figure therefore illustrates that in the absence of content validity generalisation from evidence of construct validity is problematical. As the items within a dimension are correlated (in psychometrics, by definition), achieving content validity is likely to conflict with the requirement of MAU theory that attributes should be orthogonal. In the previous analogy arithmetic and mathematical skills are likely to correlate, but imperfectly. An instrument measuring mathematical ability could not omit arithmetic (content validity), but its inclusion would cause structural dependence.

The problem is illustrated in Figure 5. Two concepts are illustrated by bolded circles, and items – linguistic statements – by rectangles. Reflecting the imprecision of language, no one item exactly corresponds with a concept. Items 1-4 are required to measure construct 1 and items 5-7 to measure construct 2. Factor analysis may be used to obtain the most efficient set of items and to omit items which cross-load (item 8). Confirmatory factor analysis or SEM may be used to achieve this goal while forcing the retention of theoretically desired constructs. However the resulting instrument structure achieves content validity by violating the DA requirement of item orthogonality which is needed to avoid double counting in utilities.

The trade-off between the psychometric requirement of content validity and the DA requirement of orthogonality has received little attention. Dimensions such as mental health are described in disease specific instruments using multiple items each of which have been shown to contribute to descriptive content. Their omission makes content validity problematical; their inclusion makes the estimation of utility preference scores analytically difficult.

Methods for the construction of instruments with content validity are described in basic psychometrics texts (Streiner and Norman 2003) and in economics by Sintonen (1994a). Sintonen, like Kaplan, argues for the use of factor analysis to reduce the number of items in a dimension but to determine the dimensions as this is an independent theoretical (and social) issue. However the techniques of SEM permit item reduction subject to the retention of nominated dimensions.

Figure 5 Construct and item overlap



Notes:
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, mental health)
Construct = A mini theory or created construct to explain observed behaviour.

Tests of content validity are possible but few have been reported. The most common have been comparisons of ceiling effects. (Predicting best health when other instruments detect poorer health indicates insensitivity – content invalidity – in this range.) As discussed earlier ceiling effects vary significantly between instruments. The Australian five instrument study tested content using overall content based upon the independent measurement of 'self TTO' as the criterion variable, ie the reduction in life people would accept for perfect quality of life. The test was two-fold, viz, to determine: (i) which instrument explained most variation in self TTO; and (ii) which instrument best explained what other instruments failed to explain, ie the residual from the first stage analysis. Results indicated a clear and similar performance of instruments on both tests, viz, 15D (greatest explanatory power) followed by SF-6D, AQoL-4D, HUI and EQ-5D (least). Each of the first four instruments explained more of the residual from the EQ-5D equation than the EQ-5D explained in the first stage.

In another test, individual attributes (content) of the EQ-5D, HUI 2 and SF-36 were each predicted from the scores obtained on the other two instruments using data from 264 German patients. Adjusted R^2 were between 0.01 and 0.57 and the authors conclude that the instrument content differs 'so much that ... (they) would produce different valuations even if other components of the instruments were the same.' (Konerding, Moock et al. 2009 p1249)

Rather than demonstrate differences, the authors of the recent Australian study reported in Figure 3 attempted to identify missing content (Khan and Richardson 2011). Instruments on the vertical axis in Figure 3 which are *relatively* sensitive to a particular dimension will have lower scores than predicted. Points will be below the line. The ratio of dimension scores of points above to below the line therefore indicates the *relative* sensitivity of the instrument. Results are shown in Table 8. Random variation generates a positive ratio so results are presented as deviations from the average ratio across the 12 possible pair-wise comparisons. As expected from Table 2, HUI has less content than other MAUI in the domains of mental health and relationships and AQoL greater content for all of the mental and social dimensions. EQ-5D is relatively sensitive to pain. Unexpectedly HUI is not significantly more sensitive with respect to senses but this is probably because the sample was small (n = 158) and there were few respondents with physical impairment.

Predictive Validity: The 'acid test' of an MAUI is whether or not it produces values with the properties required for economic evaluation. (The requirement has been described as 'Empirical Validity' (Brazier, Dolan et al. 2006) although this term is used in psychometrics to refer to all of the validation texts which draw upon empirical evidence). The properties needed for constructing valid QALYs are exacting. Since QALYs = (Life Years)(utility index) the same gain is obtained from an x percent increase in life years and an x percent increase in the utility index. This is only achieved with validity of the utility index is constructed appropriately – the 'strong interval' property (Richardson 1994). There have been few tests of this property.

Some suggest willingness to pay as a criterion for evaluating MAUIs. [link Donaldson] However the technique is controversial in the context of QALYs and no one has adopted the suggestion empirically.

As noted, several studies have employed the 'self-'TO'. In principle, with complete information, empathy and honesty this would be identical to the conventional TTO, ie people asked to imagine themselves in a health state would give the same answer as those in the health state. One study in Finland obtained average values for the self TTO which were not statistically different from 15D. However excluding the 20-59 percent of the groups who refused to trade (whose QoL was nevertheless poor) reduced the self TTO by 20 percent and the ICC between mean values to 0.58 (Honkalampi and Sintonen 2010). Two earlier studies by Stavem using self TTO and self SG in comparisons with the 15D and EQ-5D found significant differences in median scores suggesting that self referential measures invalidate rather than validate other MAUI (Stavem 1998; Stavem 1999). This conclusion is reinforced by the results of a recent five instrument Finnish study summarised in Figure 5. However the properties of self referential measures have not been discussed in the economics literature and interpreting these results is difficult.

A weak test of preferences is to determine whether most people agree that improvement has occurred when MAUI scores increase. Applying this test Roberts and Dolan (2004) found that a 0.20 increase in the EQ-5D score was necessary before 70 percent of respondents agreed that any improvement had occurred.

The logic of the Roberts-Dolan test was to use MAUI scores to test predictive validity – what people would choose. Similar logic was used earlier in a study by Nord and Richardson et al. (1993) drawing upon the identity that QALYs are the product of utility, LY and the number of people affected. From this MAUI scores were used to predict the number of people moving from a health state to full health which would be equivalent to saving one life. Results from the QWB and HUI 1 were so implausible that a survey of a population agreement would have been superfluous.

An objection to the method is that evolving multiple beneficiaries introduced an element of equity which may (or may not) have invalidated results. However the method could have been applied at the individual level as illustrated in Table 9. As with Roberts-Dolan (dis)agreement with the implications could be obtained independently from the population. Since the test is simply applied to numerous health states it is a potentially powerful and rigorous test of economic validity.

7 Conclusions

Numerous questions have been outside the scope of this review. Foremost is agreement about what is to be measured. 'Health' like 'beauty' is a vague concept and has been operationalized very differently. In effect, each MAUI has provided its own unique definition which has generally been unchallenged. The chief decision concerns the breadth and content of the definition. If an MAUI is intended strictly for use within an NHS, the definition may remain narrow, possibly excluding items extraneous to NHS funding, for example social or dental specific dimensions. The values permeating orthodox economics would suggest a broader, all encompassing approach. Anything effecting preferences should be included.

Other omitted issues include perspective and the concept of utility. Present MAUIs seek to measure personal, not social, preferences; preferences are measured as decision, not experience, utility (SWB). Challenges include a proper demonstration that MAUI have construct validity in different disease areas and, more fundamentally, economic validity; that they have the 'strong interval' property required for construction of valid QALYs.

The review has focused upon construction and validity of MAUIs narrowly defined. Scores obtained from the different MAUIs differ significantly and, consequently, QALY values and CUA ratios and the likelihood of health service funding are all significantly affected by the choice of instrument.

The numerical values obtained by MAUI's depend upon the validity of the descriptive system, the combination algorithm (model), and the scaling instrument. Of these, the evidence suggests greatest agreement between the scaling instruments. TTO, SG and even VAS values correlate fairly well. Despite this, the focus in the economics literature has been upon this choice, with overall instrument validity often judged primarily on the basis of the scaling instrument.

There has been little empirical evidence published with respect to the choice of model. Authors of the HUI and AQoL both found additive less satisfactory than multiplicative models. However these models permit double counting of content. Econometric linear models are flexible and ensure predicted values within the range of holistic utilities elicited from the sample population. But extrapolation to other populations with an additive model is problematical. Apart from AQoL-8D there has been little experimentation with non-linear models.

Least agreement exists between the items of the MAUI's descriptive system.

None of the numerous comparative studies between MAUIs and disease specific or other MAU instruments have ever concluded that a scale was invalid: but the scales differ significantly, indicating that the comparisons represent weak evidence of validity in the sense required by economic evaluation studies. Scales have been determined using different approaches and generally with little regard for content validity. It is therefore at this level that differences between instrument scores will most probably be found.

The outcome of an evaluation may presently depend upon the choice of instrument. The approach to this problem by the UK National Institute of Health and Clinical Excellence (NICE) has been to nominate a common instrument for use in all evaluations. This is the same approach as was adopted by the State of Indiana in 1897 when it sought to overcome inconsistency in the estimated values of Pi used by different bodies by attempting to legislate its value. (The Bill was rejected in the Senate!) However if the nominated choice is wrong then more harm id done than good. More realistically, instruments are neither right nor wrong. The present evidence suggests that they are more or less sensitive in different contexts. Use of a single instrument will favour interventions affecting health states where the instrument is sensitive (and the intervention efficacious) and disadvantage interventions where sensitivity is low. This indicates the need for a significant research program to determine which instruments should be used in which contexts and how to compare their values.

References

Andresen, E. M., B. M. Rothenberg, et al. (1998). "Performance of a self-administered mailed version of the Quality of Well-being (QWB-SA) questionnaire among older adults." Medical Care **36**: 1349-1360.

Belgium (2008). The Draft Pharmacoeconomic Belgian Guidelines. The Center of Expertise (KCE), INAMI/RIZIV.

Boyle, M., G. Torrance, et al. (1983). "Economic evaluation of neonatal intensive care of very low birth weight infants." New England Journal of Medicine **308**(22): 1330-1337.

Brazier, J., P. Dolan, et al. (2006). "Does the whole equal the sum of the parts? Patient-assigned utility scores for IBS-related health states and profiles." Health Economics **15**: 543-551.

Brazier, J., J. Roberts, et al. (2002). "The estimation of a preference-based measure of health from the SF-36." Journal of Health Economics **21**: 271-292.

Brazier, J., J. Roberts, et al. (2004). "A comparison of the EQ-5D and Sf-6D across seven patient groups." Health Economics **13**: 873-884.

Brooks, R. and EuroQol Group (1996). "EuroQoL: the current state of play." Health Policy **37**: 53-72.

Canada (2006). Guidelines for the Economic Evaluation of Health Care Technologies, Third Edition. Canadian Agency for Drugs and Technologies in Health (CADTH).

Capri, S., A. Ceci, et al. (2001). "Guidelines for Economic Evaluations in Italy: Recommendations from the Italian Group of Pharmacoeconomic Studies." Drug Information Journal **35**(1): 189-201.

Dolan, P. (1997). "Modeling Valuations for EuroQol Health States." Medical Care **35**(11): 1095-1108.

Dolan, P., C. Gudex, et al. (1995). A social Tariff for EuroQol: Results from a UK General Population Survey, CHE Discussion Paper 138. York, Centre for Health Economics, University of York.

Elsworth, G., J. Richardson, et al. (2011). Increasing the Sensitivity of a Quality of Life Inventory for Evaluation of Interventions Affecting Mental Health, Research Paper 61. Melbourne, Centre for Health Economics, Monash University.

EuroQol Group (1990). "EuroQol - a new facility for the measurement of health-related quality of life." *Health Policy* **16**: 199-208.

Fanshel, S. and J. Bush (1970). "A Health Status Index and its Application to Health Service Outcomes " Operations Research **18**: 1021-1066.

Feeny, D. (2002). Health-status classification systems for summary measures of population health. Summary Measures of Population Health. C. J. L. Murray, J. A. Salomon, C. D. Mathers and A. D. Lopez. Geneva, World Health Organzation**:** 329-341.

Feeny, D., W. Furlong, et al. (2002). "Multi attribute and single attribute utility functions for the Health Utilities Index Mark 3 System." Medical Care **40**(2): 113-128.

France (2004). French Guidelines for the Economic Evaluation of Health Care Technologies. The members of the Collège des Économistes de la Santé (the French Health Economists Association)

Fryback, D. G., M. Palta, et al. (2010). "Comparison of 5 health related quality of life indexes using item response theory analysis." Medical Decision Making **30**(1): 5-15.

Hawthorne, G., J. Richardson, et al. (2001). "A comparison of the assessment of quality of life (AQoL) with four other generic utility instruments." Annals of Medicine **33**(5): 358-370.

Hawthorne, G., J. Richardson, et al. (1997). The Australian Quality of Life AQoL Instrument, Working Paper 66. Melbourne, Centre for Health Program Evaluation.

Honkalampi, T. and H. Sintonen (2010). "Do the 15D scores and time trade-off (TTO) values of hospital patients' own health agree?" International Journal of Technology Assessment in Health Care **26**(1): 117-123.

Ireland (2010). Guidelines for the Economic Evaluation of Health Technologies in Ireland. Health Information and Quality Authority.

Kaplan, R., J. Bush, et al. (1976). "Health status: Types of validity and the index of wellbeing." Health Services Research **11**(4): 478-507.

Kaplan, R. M. (2005). Measuring quality of life for policy analysis: Past, present, and future. Advancing Health Outcomes Research Methods and Clinical Applications. W. R. Lenderking and D. A. Revicki. McLean VA, Degnon Associates**:** 1-35.

Kaplan, R. M., T. G. Ganiats, et al. (1998). "The Quality of Well-Being Scale: critical similarities and differences with SF-36." International Journal for Quality in Health Care **10**(6): 509-520.

Kaplan, R. M., S. Tally, et al. (2010). "Five preference based indexes in cataract and heart failure patients were not equally responsive to change." Journal of Clinical Epidemiology **doi:10.1016/j.jclinepi.2010.04.010**: Reprinted with permission from Elsevier.

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.

Konerding, U., J. Moock, et al. (2009). "The classification systems of the EQ-5D, the HUI II and the SF-6D: what do they have in common?" Quality of Life Research **18**: 1249-1261.

McCabe, C., J. Brazier, et al. (2006). "Using rank data to estimate health state utility models." Journal of Health Economics **25**(3): 418-431.

McDonough, C. M., M. R. Grove, et al. (2005). "Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among Spine Patient Outcomes Research Trial (SPORT) participants." Quality of Life Research **14**: 1321-1332.

Misajon, R., G. Hawthorne, et al. (2005). "Vision and quality of life: The development of a utility measure." Investigative Ophthalmology & Visual Science **46**(11): 4007-4015.

Mook, J. and T. Kohlmann (2008). "Comparing preference-based quality-of-life measures: results from rehabilitation patiens with musculoskeletal, cardiovascular, or psychosomatic disorders." Quality of Life Research **17**: 485-495.

Netherlands (2006). Guidelines for Pharmacoeconomic Research in the Netherlands (Updated Version, April 2006). College voor zorgverzekeringen, Diemen.

NICE (2008). Guide to the Methods of Technology Appraisal. National Institute for Health and Clinical Excellence, London.

Nord, E., J. Richardson, et al. (1993). "Social evaluation of health care versus personal evaluation of health states: evidence on the validity of four health state scaling instruments using Norwegian and Australian surveys." International Journal of Technology Assessment in Health Care **9**: 463-478.

Orlewska, E. and P. Mierzejewski. Polish Guidelines for Conducting Pharmacoeconomic Evaluations.

Orlewska, E. and P. Mierzejewski (2003). "Polish Guidelines for Conducting Pharmacoeconomic Evaluations." European Journal of Health Economics **4**(4): 296-303.

Patrick, D. L., J. W. Bush, et al. (1973). "Methods for Measuring Levels of Wellbeing for a Health Status Index." Health Services Research **8**: 228-245.

PBAC (2008). Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.3). Pharmaceutical Benefits Advisory Committee.

Peacock, S., R. Misajon, et al. (2008). "Vision and quality of life: development of methods for the VisQoL vision related utility instrument." Ophthalmic Epidemiology **15**: 218-223.

PHARMAC (2007). Prescription for Pharmacoeconomic Analysis - Methods for Cost-utility Analysis (May 2007). PHARMAC, the Pharmaceutical Management Agency.

Quercioli, C., G. Messina, et al. (2009). "Importance of sociodemographic and morbidity aspects in measuring health-related quality of life: performances of three tools " European Journal of Health Economics **10**(4): 389-397.

Richardson, J. (1994). "Cost utility analysis: What should be measured." Social Science & Medicine **39**(1): 7-21.

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, et al. (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. and M. A. Khan (2009). Preliminary results for the Validation of the Assessment of Quality of Life AQoL-8D Instrument, Research Paper 47. Melbourne, Centre for Health Economics, Monash University.

Roberts, J. and P. Dolan (2004). "To what extent do people prefer health states with higher values? A note on evidence from the EQ-5D valuation set." Health Economics **13**: 733-737.

Rosser, R. and P. Kind (1978). "A scale of valuations of states of illness: is there a social consensus?" International Journal of Epidemiology **7**(4): 347-358.

Rosser, R. M. and V. C. Watts (1972). "The measurement of hospital output." International Journal of Epidemiology **1**(4): 361-368.

Scotland (2007). Guidance to Manufacturers for Completion of New Product Assessment Form (NPAF) (Revised June 2007). Scottish Medicines Consortium.

Seymour, J., P. McNamee, et al. (2010). "Shedding new light onto the ceiling and floor? A quantile regression approach to compare EQ-5D and SF-6D responses." Health Economics **19**: 683-696.

Shaw, J., J. Johnson, et al. (2005). "US valuation of the EQ-5D health states: development and testing of the DI model." Medical Care **43**: 203-220.

Sintonen, H. (1994a). The 15-D Measure of Health-Related Quality of Life: Reliability, Validity and Sensitivity of its Health State Descriptive System, Working Paper 41. Melbourne, Centre for Health Program Evaluation, Monash University.

Sintonen, H. (1994b). The 15D-measure of health-related quality of life. II Feasibility, reliability and validity of its valuation system. Melbourne, Centre for Health Program Evaluation, Monash University.

Sintonen, H. and M. Pekurinen (1989). "A generic 15 dimensional measure of health-related quality of life (15D)." Journal of Social Medicine **26**: 85-96.

Sintonen, H., T. Weijnen, et al. (2003). Comparison of E-5D VAS valuations: analysis of background variables. The Measurement and Valuation of Health Status using EQ-5D: A European Perspecive. Evidence from the EuroQol BIOMED Research Programme. R. Brooks, R. Rabin and F. De Charro. Dordrecht, Kluwer**:** 81-101.

Stavem, K. (1998). "Quality of life in epilepsy: comparison of four preference measures." Epilepsy Research **29**: 210-209.

Stavem, K. (1999). "Reliability, validity and responsiveness of two multi attribute utility measures in patients with chronic obstructive pulmonary disease." Quality of Life Research **8**: 45-54.

Stavem, K., S. S. Frøland, et al. (2005). "Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS." Quality of Life Research **14**: 971-980.

Stewart, A., J. E. J. Ware, et al. (1977). "The meaning of health: understanding functional limitations." Medical Care **15**(11): 939-952.

Streiner, D. and G. Norman (2003). Selecting the items. Health Measurement Scales: A Practical Guide to their Development and Use Oxford, Oxford University Press**:** Ch 5.

Sweden (2003). General Guidelines for Economic Evaluations from the Pharmaceutical Benefits Board. Pharmaceutical Benefits Board (LFN).

Szende, A., Z. Mogyorósy, et al. (2002). "Methodological Guidelines for Conducting Economic Evaluation of Healthcare Interventions in Hungary: A Hungarian Proposal for Methodology Standards." European Journal of Health Economics **3**(3): 196-206.

Torrance, G. (1986). "Measurement of health state utilities for economic appraisal: A review." Journal of Health Economics **5**(1): 1-30.

Torrance, G., M. Boyle, et al. (1982). "Application of multi attribute utility theory to measure social preference for health states." Operations Research **30**(6): 1043-1069.

Torrance, G., D. Feeny, et al. (1996). "Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2." Medical Care **34**(7): 702-722.

USA (2009). The AMCP Format for Formulary Submissions (Version 3.0, October 2009). Developed by the FMCP Format Executive Committee.

Web of Science (2011). ISI Web of Knowledge, Thomson Reuters.

WHO (1948). Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June 1946, and entered into force on 7 April 1948.