MONASH University
Business and Economics

Centre for Health Economics

Research Paper 2012 (76)

Cross-national comparison of twelve quality of life instruments

MIC Paper 1
Background, questions, instruments

Professor Jeff Richardson
Position title, Centre for Health Economics
Monash University

Angelo Iezzi
Research Fellow, Centre for Health Economics
Monash University

Aimee Maxwell
Research Fellow, Centre for Health Economics
Monash University

September 2012

Centre for Health Economics
ISSN 1833-1173
ISBN 1 921187 75 1

The Multi Instrument Comparison (MIC) survey is a project funded by a National Health and Medical Research Council (NHMRC) project grant (ID 1006334) 'A cross national comparison of eight generic quality of life instruments'. (Since its inception, three additional instruments have been added).

Chief Investigators on the project are:

Professor Jeff Richardson (Monash University, Australia)
Professor Robert Cummins (Deakin University, Australia)
Professor Robert Kaplan (University of California Los Angeles, USA)
Professor Jan Abel Olsen (University of Tromso, Norway)
Professor Michael Schlander (University of Heidelberg, Germany)
Professor Joanna Coast (University of Birmingham, England)

Upon completion of the project all data will be made publicly available.

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
Jeffrey.Richardson@monash.edu

ABSTRACT

The Multi Instrument Comparison (MIC) project is the largest comparative study of health and wellbeing instruments undertaken worldwide. In total 8,817 individuals have completed twelve instruments relating to their health or wellbeing. Data were collected from a representative healthy cohort and from patients in eight clinical areas in each of five countries.

This paper presents the background to the survey, the reasons for undertaking it and the hypotheses to be tested.

Contents of the MIC survey are summarised in Box 1. Greater detail may be found in the User Manual, available from the authors.

Box 1 Country and disease area summary as at May 2012

Respondent numbers after editing (ex Norway)

Total sample	
Australia	1436
UK	1358
USA	1467
Canada	1335
Norway	unavailable
Total	5590

Health state	
Arthritis	640
Asthma	579
Cancer	577
COPD	66
Depression	617
Diabetes	641
Chronic heart disease	640
Hearing problems	595
Stroke	23
Healthy	1,212

Box 2 Main questionnaire

Type	Title	Questions
Subjective Wellbeing (SWB)	Personal Wellbeing Index (PWI)	9
	Integrated Household Survey (IHS)	5
	Satisfaction with Life Survey (SWLS)	4
	Subtotal	**18**
Multi Attribute Utility (MAU) Instruments	EQ-5D	5
	AQoL-8D and AQoL-4D	44
	HUI3	8
	15D	15
	QWB-^{SA}	77
Non-Utility	SF-36	36
	Self TTO	1
	ICECAP-A	5
Demographics		18
	Total items in composite instrument	**227**

TABLE OF CONTENTS

List of Tables

List of Figures

List of Boxes

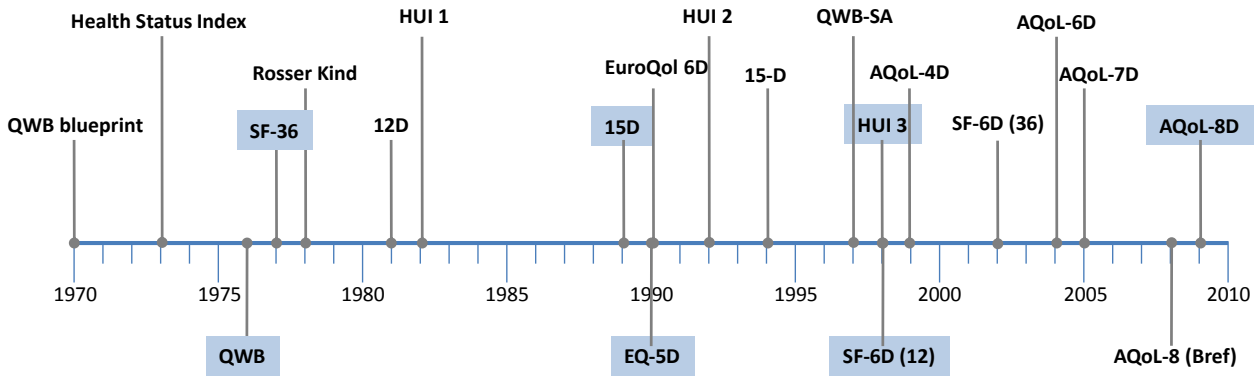
Cross-national comparison of twelve quality of life instruments

1 Introduction

Quantifying the intrinsic value of health states has proved to be conceptually and empirically problematical. 'Value' is a normative concept and its meaning has varied. Early human capital theorists calculated the value of human life as its net present value (NPV) based upon the market value of future production. The value of a health state depended upon how it affected NPV. Quality of life (QoL) was acknowledged (but categorised and largely forgotten) as an 'intangible'. Subsequently economists equated 'value' with 'utility' and the value of time in a health state was measured as quality adjusted life years (QALYs).

For a period, the chief interest in economic journals was the measurement of utility and which of four (main) measures was most appropriate – the standard gamble (SG), time trade-off (TTO), rating scale (RS) or person trade-off (PTO). The issue has never been satisfactorily resolved (Smith, Brown et al. 2008). Rather, the focus of interest has shifted. Utility is now generally measured using a multi attribute utility (MAU) instrument and multiple versions of these have emerged (see Figure 1).

Figure 1 History of MAU instruments



Source: (Richardson, McKie et al. 2011)

While economists have assumed 'value' meant 'utility' – the strength of a person's preference – psychologists have generally equated 'value' with subjective wellbeing (SWB) or happiness and there is now a large theoretical and empirical literature relating to its measurement (Kahneman, Diener et al. 2003; Helliwell, Layard et al. 2012). However, like 'utility', multiple definitions of subjective wellbeing exist. Its use in economics is reviewed by Frey and Stutzer (2002), Di Tella (2007) and Clarke et al. (2008). It has recently been suggested in the economics literature that happiness should replace utility as the definition of 'value' (Layard 2005; Dolan 2008; Dolan and Kahneman 2008).

A third concept of value has also been suggested. Drawing upon the work of Sen (1993) and Nussbaum (2000), Robeyns (2005) and Grewal et al. (2006) have developed measures of a person's basic capabilities. In principle these are the attributes which determine what an individual *can* do or experience, rather than what they *actually* do or experience.

Despite the obvious importance of the conceptualisation and measurement of 'value' – it is what we seek to achieve – there has been a remarkable dearth of comprehensive, comparative studies. Economists and psychologists largely ignored each other until the works of Kahneman and Tversky introduced elements of cognitive psychology into economics (Laibson and Zeckhauser 1998). During the early development of MAU instruments even the cross-referencing of different instruments was rare. Remarkably, there have still been relatively few thorough comparisons of instruments to determine their strengths and weaknesses and choice of instrument appears to be determined primarily by the geographic location of researchers and by history. (For a comprehensive review of MAU instruments see Richardson et al. (2011)). This lack of comparative data was the principle motivation for the present Multi Instrument Comparison (MIC) project.

2 The Problem

The fundamental problem motivating the MIC survey is that different MAU instruments produce different numbers from the same individual. This raises two questions: why does this occur and which instrument is most appropriate in a particular context. The MIC project set out to document the first problem and provide assistance in answering the second.

Empirical Evidence: In their review of empirical studies from 2005-2010 Richardson et al. (2011) identified 392 pairwise comparisons of MAU instruments. Most were the result of the inclusion of two MAU instruments in a single study and, to date, only two large and two smaller studies have included five MAU instruments simultaneously.

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 between 41-59 percent, leaving an average of 44 percent of the variance unexplained. The highest explanatory power was achieved by 15D followed by AQoL-4D (Table 1). In a more recent US study 3,844 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. Overall, 53 percent of instrument variance was unexplained (Table 1).

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 concluded that different measures give different utility values (Stavem, Frøland et al. 2005). In the context of spinal patients, lower correlations were found between EQ-5D, SF-6D, HUI 3 and QWB than in the USA and the

authors concluded 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 concluded that the instrument values were not equivalent (and) may have considerable effects upon health economic evaluation studies (Mook and Kohlmann 2008). Results of an analysis of 1,011 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).

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

Australia[(1)]	**15D**	**EQ5D**	**HUI 3**	**SF-6D**	**AQoL-4D**	**Average**
15D	1.00	0.58	0.55	0.59	0.64	
EQ-5D		1.00	0.41	0.56	0.53	
HUI 3			1.00	0.44	0.55	
SF-6D				1.00	0.55	
MEAN	0.59	0.52	0.49	0.53	0.57	0.56
USA[(2)]	**QWB SA**	**EQ5D**	**HUI 3**	**SF6D**		
QWB SA	1.00	0.41	0.45	0.43		
EQ5D		1.00	0.49	0.50		
HUI 3			1.00	0.52		
SF6D				1.00		
MEAN	0.43	0.47	0.49	0.48		0.47

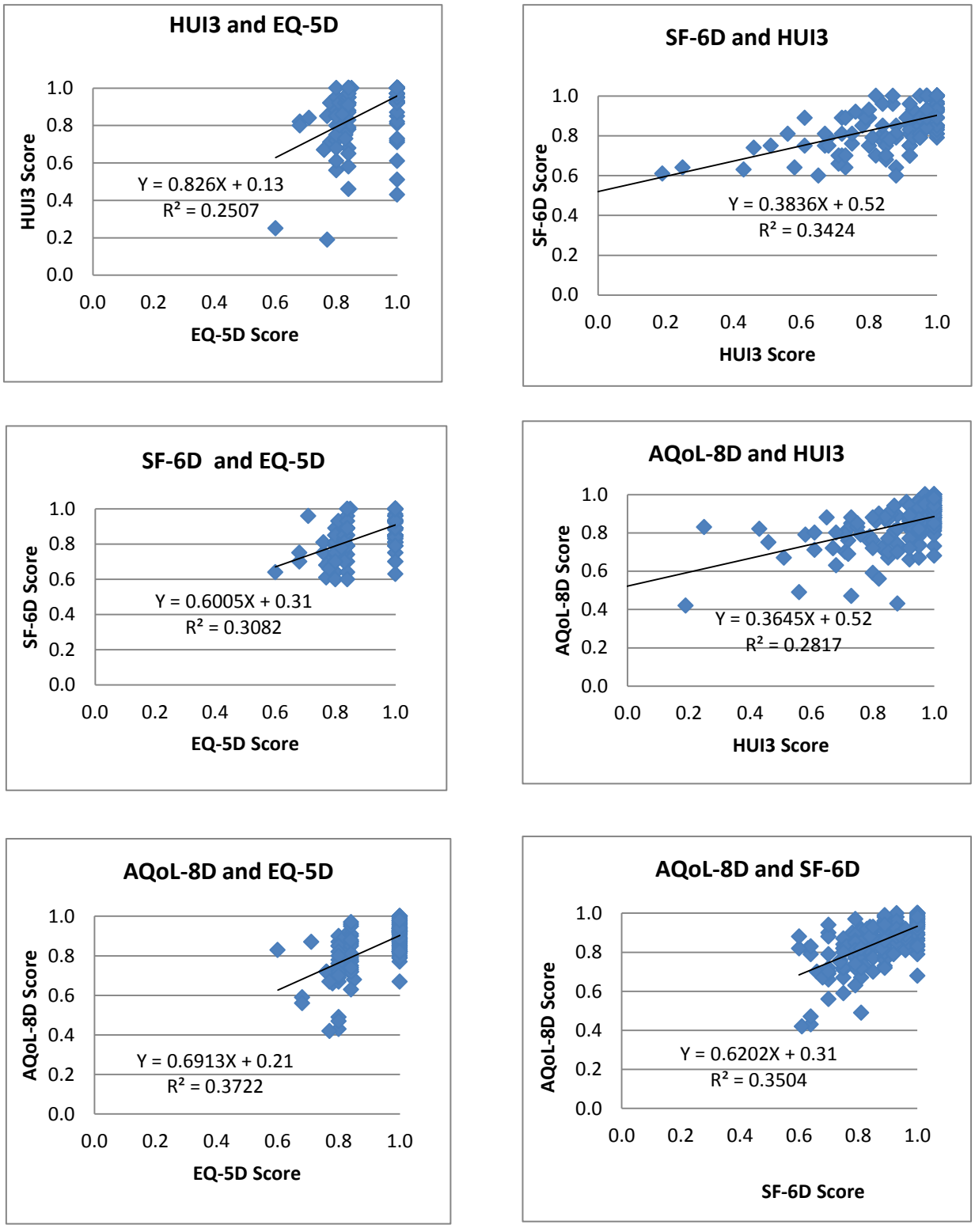
Source: [(1)]Hawthorne & Richardson (2001); [(2)]Fryback, Palta et al. (2010).

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'.

Figures 2 and 3 illustrate the discrepancy between instrument scores using results from a recent Australian study (Khan and Richardson 2009). The data reflect the strong ceiling effect of the EQ-5D (Figure 3) and the significant, but weaker ceiling effect of the HUI 3. The SF-6D and EQ-5D have the strongest floor effects with no values below 0.6 (Figure 2). Additionally, at all levels of one instrument there was significant variation in the scores predicated by other instruments. 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.

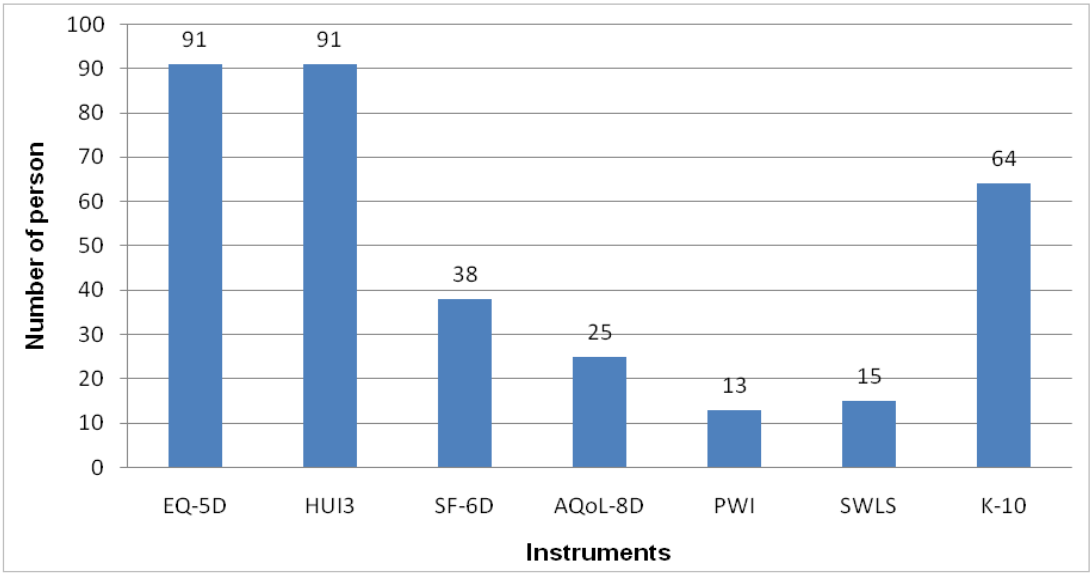
Proximate reasons: MAU instruments have two components: a descriptive system and a scoring formula. A large number of studies have been carried out to derive alternative scoring formula or algorithm, and particularly for the EQ-5D. In contrast, there has been relatively little work investigating the consistency of the descriptive systems and, in particular, their content – the concepts they measure and how completely the nuances of the concepts are conveyed by the questions in the instrument's descriptive system. The comparison of items included in different instruments discussed in the next section gives strong *prima facie* grounds for postulating that the primary cause of differences between instruments is the difference in instrument content.

Figure 2 Pairwise comparison of 4 MAU instruments



Source: Khan and Richardson (2011)

Figure 3 Ceiling effects: Numbers of individuals responding with the maximum score (non-weighted score) (n = 158)



Source: Khan and Richardson (2009)

Table 2 Dimension sensitivity: Ratio of scores: top 50% (T) to bottom 50% (B) for 4 MAU instruments

Dimensions and instruments	AQoL-8D (T-B)/se	EQ-5D (T-B)/se	HUI 3 (T-B)/se	SF-6D (T-B)/se
Dim IL	8.13	4.61	8.04	7.46
Dim Hap	13.30	8.62	9.79	7.89
Dim MH	15.02	8.21	9.54	7.64
Dim Cop	11.10	7.65	7.96	8.67
Dim Rel	14.11	7.38	8.90	8.40
Dim SW	12.60	7.43	10.62	9.16
Dim Pain	11.99	10.74	14.24	9.09
Dim Senses	12.17	5.69	11.20	9.09
K-10	13.47	10.21	10.08	11.62
PWI Score	9.48	7.37	10.72	9.80
SWLS Score	7.73	5.96	8.36	6.88

Key

ID = Independent Living; Hap = Happiness; MH = Mental Health; Cop = Coping; Rel = Relationships; SW = Self-worth; Pain = Pain; Senses = Senses; K10 = Depression Scale; PWI = Personal Wellbeing Index; SWLS = Satisfaction with Life Scale.

This hypothesis was explored in the same Australian study by exploiting the fact that, in addition to the overall MAU scores, it was possible to calculate the scores for each of the 8 dimensions of the AQoL-8D. Using each instrument, respondents were ranked, then divided into a 'top' and 'bottom' group according to their score on each instrument. The sensitivity of instruments to dimensions was tested by taking the difference in the dimension score between the two groups and standardising it with the standard error of the dimension.

The results which are shown in Table 2 indicate significant difference in the sensitivity of the EQ-5D, HUI 3 and SF-6D. Differences in the dimension scores varied from 13 to 97 percent with an average difference of 43 percent. Overall EQ-5D displays the least sensitivity with an average standardised difference of 7.54 and HUI 3 shows the greatest sensitivity with an average standardised difference of 10.0.

These results are preliminary as they were derived from a small sample of 158 healthy individuals (Khan and Richardson 2009). Differences between patients in severe health states might, potentially, be much greater.

3 MIC objectives

3.1 Documenting differences

Previous literature indicates major discrepancies in the scores produced by different instruments. The first objective of the MIC project was to document these differences using a large database and to determine the extent of the problem in different disease areas and different levels of disease severity. Differences between countries may also be examined.

3.2 Validity

Convergent validity – the correspondence of scores with the scores of other instruments purporting to measure the same quantity – may be examined at least three ways. First, comparison of the MAU instruments provides a form of 'cross validation'. Secondly MAU instrument scores may be compared with five other instrument scores with which they would be expected to correlate. These are results from the SF-36, the three subjective wellbeing (SWB) instruments and from the self TTO instrument described in the next section. Thirdly, the MAU instruments may be compared with the disease-specific instruments included in the survey.

Predictive validity may be tested by a simple extension of the previous analysis. When independent or other MAU instrument scores are shifted from low to high values an MAU instrument should predict a corresponding change. Sensitivity to change may vary with the severity of the health states and this will be examined.

3.3 Content analysis

Each instrument consists of a unique set of questions. Differences in the literal meaning and in the interpretation of these implies the likelihood of significant differences in the sensitivity of instruments to different dimensions of health. The MIC data allow a detailed examination of this as the SF-36, AQoL-8D, and PWI provide a total of 25 (overlapping) independently validated dimensions of health covering subjective wellbeing, physical, mental and social handicap. Instrument content may be examined by a comparison of instruments with these dimensions using both regression and psychometric factor analysis. Pairwise comparisons of instruments may be conducted.

3.4 Proximate cause of differences

MAU instruments have two components: a descriptive system – the subject of the comments above – and a utility scoring formula which converts item responses into a single utility score. There is a long tradition in psychometrics of eschewing the use of utility or other weights and calculating scores by giving equal weighting to items and dimensions. A comparison of scores obtained in this way with the scores obtained from the utility formula permits the estimation of the respective roles of the two components, the descriptive system and the scoring formula.

3.5 Relation between conceptually dissimilar scales

While MAU instruments are the principle focus of the study, the survey contains three SWB instruments and the ICECAP Capabilities instrument. There has been no comparison of these three types of instruments in the literature and the relationship between them will be analysed using factor and multivariate analysis.

3.6 Transformations

When different instruments produce different scores, a first step towards reconciling the problem is the creation of a transformation between them. This cannot overcome instrument insensitivity but it can, minimally, harmonise the magnitude of the units. The MIC project allows this to be achieved for each of the different disease groups.

4 The Instruments

4.1 MAU instruments

The MAU instruments in the study are described in Brazier et al. (2007) and Richardson et al. (2011).

Construction of an MAU instrument 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 a formula or algorithm for extrapolating results. Table 3 summarises the MAU descriptive systems. Two broad approaches to description ('conceptual type') have been used. Following the WHO typology health problems result in impairment, disability and handicap; or, in more recent terminology, 'body function and structure', 'activity' and 'participation'. Three MAU instruments 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 MAU instruments have adopted a 'within-the-skin' approach (disability) – 15D and HUI 3 – although 15D was modified to include one handicap dimension (usual activities). The QWB^{SA} spans all concepts.

The resulting instruments have between 5 and 15 dimensions with one item per dimension in HUI 3, 15D, EQ-5D and SF-6D and an average of 4 items per dimension for AQoL-8D. QWB^{SA} has 3 basic dimensions supplemented with 35 symptom/problem groups which transcend dimensions. Item response levels in the instruments vary from 3 to 6. Until a recent revision of the response levels from 3 to 5, EQ-5D defined 243 possible health states. The 7 MAU instruments in the present study now have between 3,125 (EQ-5D) and 2.37×10^{23} (AQoL-8D) health states. Larger instruments, particularly AQoL, define numerous 'empty' states (eg 'bedridden' and 'no problems with self-care').

Dimensions overlap imperfectly (Table 4). Several are unique to a particular instrument and similarly named dimensions include different items. Consequently, to appreciate instrument content requires examination of the items. 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 (eg brevity). However, the differences are potentially important for instrument validity.

Table 3 Instrument descriptive systems

	QWB[SA]	15D	EQ-5D	HUI 3	SF-6D	AQoL-8D	SF-36
Descriptive system							
Conceptual type	Handicap Disability Impairment	Disability (handicap)	Handicap (disability)	Disability	Handicap (disability)	Handicap (disability)	Handicap
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	Psychometric reduction of MOS
Dimensions	3 + 27 symptoms/problems	15	5	8	6	8	8
Items		15	5	8	6	35	36
Response levels	2, 3 (2)	4-5	5	5-6	4-6	4-6	2-6
States defined	945	3.1×10^{10}	3,125*	972,000	18,000	2.37×10^{23}	8.7×10^{20}

*Until a recent increase in the response levels from 3 to 5 EQ-5D had 243 response levels.

Table 4 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**	**SF-36**
Physical	Physical ability/ Vitality/Coping/ Control	……..	*			*	**	******
	Bodily Function/ Self Care	…….	***	*			*	*
	Dexterity				*			*
	Pain/Discomfort	……..	*	*	*	*	**	**
	Senses	…..	**		**		**	
	Usual activities/ Work function	……...	*	*		*	****	****
	Mobility/walking	……..	*	*	*		*	***
	Communication	..	*		*		*	
	Vitality							****
Psycho-social	Sleeping	.	*				*	
	Psychological: Depression/Anxiety/ Anger	….	***	*	*	*	*******	***
	General Satisfaction						****	*
	Self Esteem						**	
	Cognition/Memory Ability	.			*			
	Social Function/ Relationships					*	******	**
	(Family) Role					*	*	***
	Intimacy/Sexual Relationships	.	*				*	
	General health							******
	Number of items		**15**	**5**	**8**	**6**	**35**	**36**

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 5 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.

Table 5 summarises the methods used to model the utility scores for each of the MAU instruments and some of the resulting utility values. Like the characteristics revealed in the previous two tables, the most notable result emerging from Table 5 is the differences, not the similarities between instruments.

4.1.1 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. While these contain no explicit 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:

$$VALUE = 1\text{-}D_1\text{-}D_2\text{-}D_3\text{-}S$$

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 MAU instrument. 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 manual may be obtained at https://hoap.ucsd.edu/qwb-info/

4.1.2 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 = 2,500) and transformations of VAS data into 56 'utilities' using an econometric transformation. Results demonstrated convergent validity of 15D scores (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

4.1.3 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. There are sixteen versions in English varying with the mode of administration, assessment viewpoint and duration of assessment period. The website is http://fhs.mcmaster.ca/hug/.

4.1.4 EQ-5D: The 5 item 5 level EQ-5D defines 3,125 health states. It was originally designed to compare broad preference patterns across Europe and not as a stand-alone MAU instrument 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 MAU instrument and eventually became the preferred instrument by the UK National Institute for Health and Clinical Excellence (NICE). The UK weights, which are the most widely used, employed VAS and TTO data from a survey of 2,997 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 to 12 years has been translated into 12 languages. An algorithm has been estimated in the USA using data from 3,773 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/.

4.1.5 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.

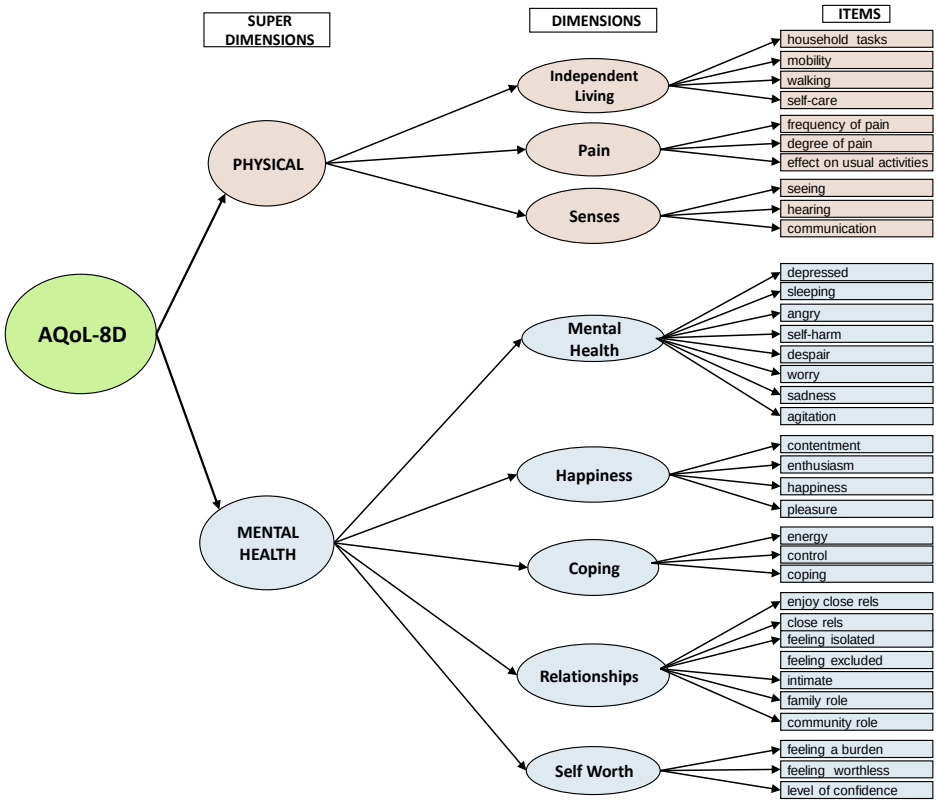
The items of the descriptive system were derived from the factor analysis and psychometric properties undertaken in developing of the SF-36.

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. A 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.

4.1.6 Assessment of Quality of Life (AQoL-8D): 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 4 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 AQoLs 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 15 item instrument scaled without the original dimension for symptoms) has been reduced to an 8 item AQoL, the AQoL-8 (which should not to be confused with the 35 item AQoL-8D). The four AQoL instruments have been translated into traditional and simplified Chinese, Spanish, German, Italian, Arabic, Norwegian and Danish. The AQoL website is http://www.aqol.com.au/.

4.2 Non MAU instruments

The contents of the SF-36 are similar to the contents of MAU instruments and are summarised in Tables 3 to 5. Elements of the three Subjective Wellbeing (SWB) instruments are summarised in Table 6. This indicates that, like MAU instruments, the content of the instruments differ significantly (despite the generic label 'SWB').

Table 6 Elements in Subjective Wellbeing (SWB) instruments

	Self-worth/ Achievement	Happiness/ Anxiety	Satisfaction	
			General	Elements: standard of living health safety security religion community
PWI	*	-	*	
SWLS	**	-	***	-
IHS	-/-	*/*	*	-

4.2.1 Short Form 36 (SF-36)

The Short Form 36 (SF-36) was developed from the 245 items of the Rand Medical Outcomes Study (Lohr, Brook et al. 1986) and is the most widely used instrument for measuring HRQoL worldwide. From the original 40 physical and mental concepts, 8 were selected using psychometric procedures. Reliability has been established on numerous occasions and in numerous countries (McHorney, Ware Jr et al. 1994; Gandek, Ware et al. 1998). Validity and sensitivity have likewise been tested in numerous contexts and countries. (For a review see McDowell 2006). Use of the SF-36 is described in its manual and explained on the SF-36 website (http://www.sf-36.org/).

4.2.2 The Personal Wellbeing Index (PWI)

The PWI scale contains 8 items relating to satisfaction. Each corresponds to a quality of life domain and the 8 domains, theoretically are the minimum set of domains that represent the first level deconstruction of the global question: 'How satisfied are you with your life as a whole?' The instrument was developed by Cummins, Eckersley, Pallant, Van Vugt, and Misajon, (2003). It is further described on its website http://www.deakin.edu.au/research/acqol/index_wellbeing/index.htm

4.2.3 Integrated Household Survey (IHS) instruments

The Integrated Household Survey (IHS) is a composite survey combining questions asked in a number of UK Office for National Statistics (ONS) social surveys (ONS 2012). The aim of the IHS is to produce estimates for particular themes to a higher level of precision and at a lower geographic level than is possible in individual ONS social surveys. The IHS includes two sections, viz, a suite of core IHS questions and individual survey modules 'bolted' onto the core.

Current modules of the IHS include the General Lifestyle Survey (GLF) which is included in the MIC survey. It consists of four satisfaction questions each with eleven response categories (0-10). Questions relate to (i) your life nowadays; (ii) things in your life being worthwhile; (iii) happiness; and (iv) anxiety.

Information on IHS may be found at http://www.ons.gov.uk/ons/guide-method/surveys/respondents/business/a-z-of-business-surveys/integrated-household-survey/index.html

4.2.4 Satisfaction with Life (SWLS)

The Satisfaction with Life Scale (SWLS) was developed by Diener et al. (1985) as a measure of life satisfaction. This is believed to be one of three factors in the more general construct of subjective wellbeing. The other two factors are positive and negative affective appraisal. Life satisfaction is distinguished from affective appraisal as it is determined primarily by cognitive rather than emotional factors. The SWLS consists of 5 items and a response scale from strongly disagree to strongly agree. The items are (i) in most ways my life is close to my ideal; (ii) the conditions of my life are excellent; (iii) I am satisfied with my life; (iv) so far I have gotten the important things I want in life; and (v) if I could change my life over I would change almost nothing. Information on the SWLS may be found at http://www.tbims.org/combi/swls/.

4.2.5 ICE-CAP Capabilities

The ICECAP is a measure of capability for use in economic evaluation (Grewal, Lewis et al. 2006). Unlike most profile measures used in economic evaluations, the ICECAP focuses on wellbeing defined in a broader sense, rather than health. The measure covers attributes of wellbeing that were found to be important to older people in the UK. Values attached to ICE-CAP do not reflect preferences. It is based upon judgements of what is (or should be) important.

ICECAP comprises five attributes (the lay terms are in brackets):

i. Attachment (love and friendship)
ii. Security (thinking about the future without concern)
iii. Role (doing things that make you feel valued)
iv. Enjoyment (enjoyment and pleasure)
v. Control (independence)

The ICECAP-O descriptive system was developed using qualitative methods. A set of index values for the ICECAP-O have been estimated using a best-worst scaling study of older people in England.

Further information can be found on the ICECAP-O website http://www.birmingham.ac.uk/research/activity/mds/projects/HaPS/HE/ICECAP/ICECAP-O/index.aspx

4.2.6 Self TTO

The Self TTO is conceptually similar to the time trade-off (TTO) as it asks individuals to consider a reduced life expectancy in perfect health in exchange for a longer life in imperfect health. The difference is that the imperfect health state of the Self TTO is not constructed to resemble the health state of interest and then presented more or less as an abstraction. Rather, respondents are told that the imperfect health state is their present health state. The trade-off presented is between a better, shorter life and a continuation exactly as they are at the time of the interview.

With perfect imagination and emotional detachment these differences in framing would be unimportant. However, previous use of the instrument indicated less willingness to trade with the Self TTO and a relatively low correlation with other instruments (Hawthorne, Richardson et al. 2003).

For the MIC project an online version of the Self TTO was created with an avatar talking subjects through the 'flip flop' set of choices which converged upon a final value. As at June 2012 the technique was undergoing validation tests which will be published as Iezzi, Heriott et al (2012). The Self TTO may be accessed and downloaded from the AQoL website http://www.aqol.com.au

Table 7 Dimensions and elements in disease instruments used in multi instrument comparison

		Mental Health		Hearing	Asthma	Diabetes	Arthritis	Heart	Cancer	MAU
Dimension	**Element**	**DASS21**	**K10**	**APHAB**	**AQLQ**	**D-39**	**AIMS2-SF**	**MACNEW**	**QLQ C-30**	**AQoL-8D**
Independent Living	jobs around the house/self care					**	*			**
	mobility /mobility outside home				**	**	******		***	**
	upper body dexterity						*****			
Pain	pain						***		**	***
Senses	following conversation/understanding			**********						
	hearing/effect of noises			********						*
	vision					*				*
Happiness	enthusiasm/initiative	**					*	*		****
Mental Health	hopelessness, despair, sad, crying	****	***		*	*	*	**	*	***
	frustration/angry/intolerance	***			*			*	*	*
	self harm									*
	worry/anxiety/tense/fear/sleep	******	**		****	***	**	**	***	**
	agitation/restlessness	*	**					*		*
Coping	energy/dizziness/coping/control		**		*	**********		***	***	***
Relationships	family/community/social activity				***	**	****	****	*****	****
	excluded							*		*
	physical restriction/dependence				*****			***		
	sexual functioning					***		*		*
Self Worth	burden					*	*	**	***	*
	self-worth/confidence	*	*			*		****		**
Physiological	sob/wheezing/tight chest	*			***			**	*	
	dry mouth/breathing/trembling	***								

		Mental Health		Hearing	Asthma	Diabetes	Arthritis	Heart	Cancer	MAU
Dimension	**Element**	**DASS21**	**K10**	**APHAB**	**AQLQ**	**D-39**	**AIMS2-SF**	**MACNEW**	**QLQ C-30**	**AQoL-8D**
	gastro intestinal								*****	
Cognitive	reading/memory								**	
Financial	financial								*	
General Health	general health					***			**	
	number of items	21	10	24	20	39	24	27	30	

4.3 Disease-specific instruments

A large number of disease-specific instruments exist for each disease state. Reviews of these are provided in Bowling (1995), McDowell (2006) and elsewhere. Instruments were selected for the MIC project on the basis of the available reviews and with the advice of Australian researchers in the different disease areas. The selected instruments are summarised in Table 7.

4.3.1 Depression

The Depression Anxiety Distress Scale (DASS21) is a short version of the DASS42 developed at the University of New South Wales with non-clinical samples (Lovibond and Lovibond 1995). The instrument is a set of three self-report scales developed to assess the severity of the core symptoms of depression, anxiety and stress. It is based on a dimensional concept of psychological disorder and is not a categorical measure of clinical diagnosis.

The Depression Scale assesses dysphoria (state of unease), hopelessness, devaluation of life, self-deprecation, lack of interest/involvement, anhedonia (inability to experience pleasure), and inertia. The Anxiety Scale assesses physiological effects, situational anxiety, and subjective experience of anxious affect. The Stress Scale assesses difficulty relaxing, nervous arousal, and being easily upset/agitated, irritable/over-reactive and impatient. Subjects are asked to use 4-point severity/frequency scales to rate the extent to which they have experienced each state *over the past week*. Scores for depression, anxiety and stress are calculated by summing the scores for the relevant items.

The 3 dimensions of the DASS21 contain 7 items each rated on a 4-point Likert scale of frequency or severity over the past week as shown in the box below.

Over the past week	
0	Did not apply to me at all
1	Applied to me to some degree, or some of the time
2	Applied to me to a considerable degree, or a good part of the time
3	Applied to me very much, or most of the time

Note that the following scores are based on the full (42 items) scores. Scores collected using the DASS21 need to be doubled to compare with this table. Scores for depression, anxiety and stress are calculated by summing the scores for the relevant items and characterise degrees of severity as shown below:

	Depression	**Anxiety**	**Stress**
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	28+	20+	34

Further information regarding this instrument can be found on the DASS website http://www2.psy.unsw.edu.au/groups/dass//

The **Kessler Psychological Distress Scale (K10)** is a short measure of non-specific psychological distress based on questions about the level of nervousness, agitation,

psychological fatigue and depression. It contains 10 items with 5 response levels each measuring frequency.

Over the past 4 weeks	
Score 5	All of the time
Score 4	Most of the time
Score 3	Some of the time
Score 2	A little of the time
Score 1	None of the time

The K-10 was developed by Prof Ronald Kessler, Harvard for use in the US National Health Interview Survey (NHIS) to discriminate between those people who have a serious mental illness and those who do not (Kessler, Barker et al. 2003). It produces a global measure of 'psycho-social distress' based on questions about the level of anxiety and depressive symptoms in the past four weeks.

Scores are summed and range from 10 to 50. Higher scores indicate greater psychological distress as shown below.

10-19	Likely to be well
20-24	Likely to have a mild depression and/or anxiety disorder
25-29	Likely to have a moderate depression and/or anxiety disorder
30-50	Likely to have a severe depression and/or anxiety disorder

People who score 0-15 comprise 78 percent of the population. They have one quarter the population risk of having an anxiety or depressive disorder and only a remote chance of reporting a suicidal attempt in their lifetime.

People who score 16-30 comprise 20 percent of the population and have a one in four chance (3 times the population risk) of having a current anxiety or depressive disorder and 1 percent chance (3 times the population risk) of ever having made a suicide attempt.

People who score 30-50 have a three out of four chance (10 times the population risk) of having an anxiety or depressive disorder and 6 percent chance (20 times the population risk) of ever having made a suicide attempt.

Further information may be found at http://www.hcp.med.harvard.edu/ncs/k6_scales.php

4.3.2 Hearing Loss

The **Abbreviated Profile of Hearing Aid Benefit (APHAB)** was developed at the Hearing Aid Research Laboratory at the University of Memphis (Cox and Alexander 1995). It is a 24 item instrument in which patients report the amounts of trouble they are having with communication or noises in various everyday situations. It contains four subscales: (i) Ease of Communication (EC), (ii) Reverberation (RV), (iii) Background Noise (BN); and (iv) Aversiveness of Sounds (AV).

The instrument was developed to measure the disability associated with hearing loss and the reduction of disability achieved with a hearing aid.

In our study the instrument is used to obtain a measure of disability on only one occasion without any follow up. Such instruments are normally used to measure the difference between unaided hearing and aided hearing once the patient has adjusted to the hearing aid (after 2 to weeks).

Each item has 14 response levels: 7 response levels for frequency of problems 'without hearing aids' and 7 response levels for frequency of problems 'with hearing aids'. For some items, an answer of 'always' indicates a few problems and generates a low score. Other items are written so that an answer of 'always' indicates a lot of problems and generates a high score.

The instrument score is the mean of the scores for all the items in the EC, RV and BN subscales as shown below.

	Not a reversed item	Reversed Item
A	Always (99%)	1%
B	Almost Always (87%)	12%
C	Generally (75%)	25%
D	Half-the-time (50%)	50%
E	Occasionally (25%)	75%
F	Seldom (12%)	87%
G	Never (1%)	99%

Further information may be found at http://www.memphis.edu/csd/harl/aphab.htm

Other instruments considered also included measuring the effect of rehabilitation:

- Hearing Handicap Inventory for Adults (HHIA)
- Client Oriented Scale of Improvement (COSI) is a clinical tool developed by the National Acoustic Laboratories (http://www.nal.gov.au/nal) for outcome measurement. It is an assessment questionnaire for clinicians to use on their patients hence not appropriate for an online survey
- Glasgow Hearing Aid Benefit Profile (GPHAB). This instrument contains 56 items.

4.3.3 Asthma

The **Asthma Quality of Life Questionnaire (AQLQ)** is a 20 item instrument developed to measure quality of life in adults with asthma (Marks, Dunn et al. 2006). Development of the questionnaire commenced with qualitative research and progressed through principal components analysis to thorough testing of the psychometric properties of the final questionnaire.

It has four subscales (i) Breathlessness; (ii) Mood disturbance; (iii) Social disruption, and (iv) Concerns for health. Scoring is additive to produce a total scale score together with subscale scores. All items have 5 response levels as follows:

Not at all	1
Mildly	2
Moderately	3
Severely	4
Very Severely	5

Further information is available from http://www.qoltech.co.uk/

4.3.4 Diabetes

The **Diabetes-39 Questionnaire (D-39)** was developed to assess the quality of life of diabetic patients (Boyer and Earp 1997). It covers five dimensions of health: (i) Energy and mobility (15 items); (ii) Diabetes control (12 items); (iii) Anxiety and worry (4 items); (iv) Social burden (5 items); and (v) Sexual functioning (3 items). Each item has 7 response levels ranging from 'Not at all affected' (score =1) to 'Extremely affected' (score = 7).

The original instrument has a rating scale between 0.5 and 7.5 with 0.5 intervals possible. The MIC study has limited the responses to whole numbers (1-7). The raw score for each scale is calculated by adding the responses for individual items. Possible ranges for raw scale scores are:

Diabetes Control	12-84
Anxiety and Worry	4-28
Social Burdening	5-35
Sexual Functioning	3-21
Energy and Mobility	15-105

Scores are additive and the raw, summated rating is transformed to a 0-100 scale using a linear transformation.

4.3.5 Arthritis

The **Arthritis Impact Measurement Scales 2 Short Form (AIMS2-SF)** consists of 26 items and was developed by Guillemin, Coste et al. (1997). The original AIMS2 contains 78 items covering physical, social and emotional wellbeing (Meenan, Mason et al. 1992). The Short Form was developed with psychometric properties similar to those of the AIMS2. It has 5 scales: (i) Physical; (ii) Symptom (pain); (iii) Affect; (iv) Social interaction; and (v) Role.

All items have 5 response levels

All days	5
Most days	4
Some days	3
Few days	2
No days	1

The score is standardized to a 0-10 scale. The total health score is calculated by summing the standardized scores.

4.3.6 Heart Disease

The **MacNew Heart Disease Health-Related Quality of Life Questionnaire (MacNew)** contains 27 items designed to evaluate physical limitations, emotional, and social functioning are affected by coronary heart disease and its treatment (Höfer, Lim et al. 2004). It is a modification of the original Quality of Life after Myocardial Infarction (QLMI) questionnaire. Items cover the previous two weeks and have 7 response levels.

Scoring for each item is 1-7 (poor HRQL to high HRQoL), and a global HRQL score can be calculated as the average over all the scored items.

4.3.7 Cancer

The **Cancer Quality of Life (QLQ) C-30**, developed by the European Organisation for Research and Treatment of Cancer (EORTC) (Aaronson, Ahmedzai et al. 1993). The instrument contains 30 items, 28 of which have the following responses

Not at all	1
A little	2
Quite a bit	3
Very much	4

The remaining two items are global self-perceived quality of life VAS questions covering the past week.

The questionnaire contains five functional scales (physical functioning, role functioning, cognitive functioning, emotional functioning, social functioning), general quality of life, three symptom scales (fatigue, nausea and vomiting, pain) and six single items.

Scales are summed to produce subscale scores.

4.3.8 COPD

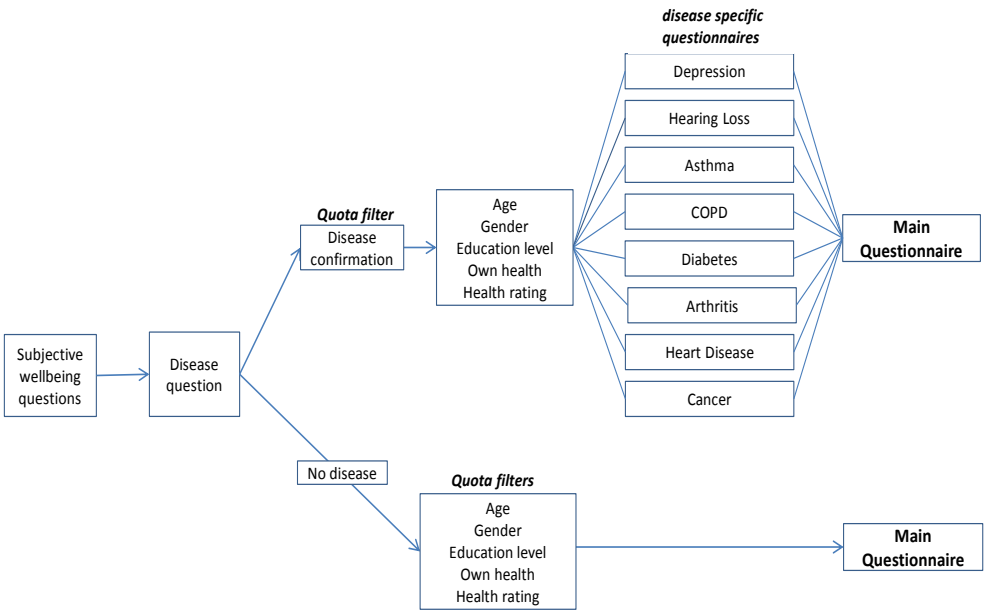
The St George's Respiratory Questionnaire for COPD Patients (SGRQ-C) is a 40-item instrument developed to measure health impairment in patients with asthma and COPD (Meguro, Barley et al. 2007). The subscales are: Symptoms, Activity, and Impacts. It is a shorter version stemming from the original SGRQ derived by removing the items with the weakest measurement properties.

The scoring algorithm calculates a total and three component scores. Owing to the poor response numbers in Australia this disease was dropped from the MIC survey.

5 Survey administration and editing

The administration of the MIC survey is illustrated in Figure 5. The survey company, CINT, invited individuals on their respondent database to participate. The introductory letter from Monash University is reproduced in Box 3. A new person accepting this invitation was first asked to complete the three subjective wellbeing questions: the Personal Wellbeing Index (PWI), the Integrated Household Survey (IHS) and the Satisfaction with Life Survey (SWLS). These questions were administered immediately as they seek to measure 'affect' – a person's 'undigested' feelings. Asking the questions after 'priming' respondents with questions about their health (do you have one of the eight diseases of interest?) would potentially create biased responses.

Figure 5 Administration of the MIC online questionnaires



After

completion of these questions the respondent was asked the following question: 'Have you got a current diagnosis of any of the following health problems? Please choose the most serious illness you have.' Depending upon their answer respondents were assigned to one of two groups – those with diseases of interest and those without. The system redirected the group with a disease of interest to a confirmation question: 'Please confirm which of these (listed illnesses) is the most serious diagnosed illness you have'. Those nominating one of the survey diseases proceeded further with the survey if and only if the quota – the target number of respondents – had not been reached. Disease group respondents then completed the core questionnaire followed by a disease-specific questionnaire.

Box 3 Introductory invitation from Monash University

International Quality of Life Measurement Study

Welcome to the International Quality of Life Measurement Study, the most comprehensive study of its kind ever undertaken. **Please complete this survey with valid and considered answers – our analyses rely on your honesty**.

The study is a collaboration between researchers in Australia, Canada, Norway, the UK and the USA. It compares questionnaires used to measure health-related quality of life around the world. The findings will help researchers and health authorities choose the best questionnaire for understanding how health services affect people's quality of life.

Some questions may seem repetitive. A proper comparison of the different questionnaires that are used worldwide requires that we ask the questions exactly as they appear in the original versions.

The survey should take around 25 minutes to complete and your answers will be confidential. Please save the Participant Information form (this link will open in a new tab).

Consent

- I have read the Online Participant Information/Explanatory Statement and I agree to take part in the Monash University research project specified.
- I understand that my participation is voluntary, and that I can withdraw at any stage of the project without being penalised or disadvantaged in any way.
- I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party.
- I understand that data from this online survey will be kept in a secure storage and accessible to the research team. I also understand that the data will be destroyed after a 5 year period unless I consent to it being used in future research.

I am satisfied with the information provided in the statement, and if I have any questions in future I am satisfied that I can contact one of the investigators mentioned in the Participant Information/Explanatory Statement (this link will open in a new tab).

When you enter the survey you are giving consent. Press the NEXT button to proceed.

Thank you for your assistance.

Yours sincerely

Professor Jeff Richardson (Research Coordinator)
Foundation Director
Centre for Health Economics
Monash University

Box 4 Introduction by CINT

> To ensure quality data, all responses are checked for validity (on multiple levels), particularly responses which are inconsistent throughout the survey. Responses which are unable to be validated may lead to incentives being withheld.

To avoid order effects in the overall results the order in which remaining questionnaires were administered was randomised by the system.

Those who did not report a disease were questioned about their age, gender and education. Additionally they were asked to indicate their overall health on a visual analogue scale (VAS) where *'zero is the least desirable state of health you could imagine and 100 is the best possible health (physical, mental and social).*' Since the objective was to ensure that the cohort had relatively good health the individual was invited to proceed to the core questions only if their VAS score exceeded 70 and their age, gender and education quota had not been filled.

5.1 Editing

Introductory comments from the panel company to their panellists were designed to deter unreliable respondents. (See Box 4). Eight edit criteria were subsequently supplied to eliminate unreliable answers. These were:

Edit 1: Any response that was completed in less than 20 minutes was eliminated. The survey median completion time was 32 minutes (range 7.7-260.9 minutes). Times between 20-25 minutes were marked for subsequent inspection (Edit 7, 8).

Edit 2: The EQ-5D mobility question was duplicated in the survey. Anyone with a response that varied by more than +/- 1.00 was eliminated. Those differing by only +/- 1.00 were earmarked for subsequent inspection (Edit 7, 8).

Edit 3: The SF-36 question 1 and question concerning own health were identical. Those with responses greater than +/- 1.00 were eliminated. Those without identical answers but within +/- 1.00 were earmarked for examination along with other edit criteria.

Edit 4: SF-36 question 1 and QWB question 9a were identical. The same criteria were applied as above.

Edit 5: Own health and QWB question 9a were identical. The same criteria were applied as above.

Edit 6: EQ-5D question 4 (pain) and AQoL-8D question 22 (pain) were very similar. Those with two response level differences were eliminated.

Edit 7: The number of inconsistencies from edits 2, 3, 4, 5, 6 were summed. Those with two or more inconsistencies and a time less than 25 minutes were eliminated.

Edit 8: Those with three or more inconsistencies were eliminated.

The effect of these procedures on the sample size of Australian respondents with self-reported disease is shown in Table 8.

Table 8 Effect of edit procedures on the sample size: Australian disease-group

Stage	Deleted	Remaining
		1376
Edit 1	36	1340
Edit 2	16	1324
Edit 3	10	1314
Edit 4	3	1311
Edit 5	7	1304
Edit 6	72	1232
Edit 7	20	1212
Edit 8	41	1171
Total deleted	205	14.9%

6 Conclusion

With an ageing population and costly technologies for improving the quality of life it is likely that there will be increasing reliance upon cost utility analysis (CUA) to determine which quality enhancing therapies should be adopted by national health schemes. But the integrity of CUA depends upon the integrity of the instruments which measure the quality of life and the present evidence indicates that neither the reliability nor validity of these instruments can be assumed.

The emphasis in the literature to date has been upon demonstrating the validity of particular instruments in limited comparisons. As none of the instruments is without merit, these studies generally succeed in validating them – the hypothesised correlation with another instrument is confirmed and the 'glass is found to be half full'. The important task of prioritising between instruments and demonstrating which are not sensitive in particular contexts, has not been rigorously undertaken. The data collected in the MIC project was designed to focus attention upon the other half of the glass and help researchers and health authorities in their choice of instrument.

The present paper has presented the background to the MIC study: the evidence for its need, its objectives, its protocol and questionnaire. Subsequent papers will present the results.

References

Aaronson, N. K., S. Ahmedzai, et al. (1993). "The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology." Journal of the National Cancer Institute **85**(5): 365-376.

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.

Bowling, A. (1995). *Measuring desease: A review of desease-specific quality of life measurement scales*. Buckingham, Philadelphia, Open University Press.

Boyer, J. G. and J. A. L. Earp (1997). "The development of an instrument for assessing the quality of life of people with diabetes." Medical Care **35**(5): 440-453.

Brazier, J., J. Ratcliffe, et al. (2007). Measuring and Valuing Health Benefits for Economic Evaluation. Oxford, Oxford University Press.

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

Clark, A., P. Frijters, et al. (2008). "Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles." Journal of Economic Literature, American Economic Association **46**(1): 95-144.

Cox, R. M. and G. C. Alexander (1995). "The abbreviated profile of hearing aid benefit." Ear and Hearing **19**(2): 176-186.

Cummins, B., R. Eckersley, et al. (2003). "Developing a national index of subjective wellbeing: the Autsralian Unity Wellbeing Index." Social Indicators Research **64**(2): 159-190.

Di Tella, R., J. Haisken-De New, et al. (2007). "Happiness Adaptation to Income and to Status in an Individual Panel." Journal of Economic Behavior & Organizatio **76**(3): 834-852.

Diener, E., R. A. Emmons, et al. (1985). "The Satisfaction with Life Scale." Journal of Personality Assessment **49**: 71-76.

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

Dolan, P. (2008). "In defence of subjective well-being." Health Economics, Policy and Law **3**: 93-95.

Dolan, P. and D. Kahneman (2008). "Interpretations of utility and their implications for the valuation of health." Economic Journal **118**(525): 215-234.

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.

Frey, B. and A. Stutzer (2002). "What can economists lean from happiness research?" Journal of Economic Literature **XL**: 402-435.

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.

Gandek, B., J. Ware, et al. (1998). "Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: results from the IQLA project." Journal of Clinical Epidemiology **51**: 1149-1158.

Grewal, I., J. Lewis, et al. (2006). "Developing attributes for a generic quality of life measure for older people: preferences or capabilities?" Social Science & Medicine **62**: 1891-1901.

Guillemin, F., J. Coste, et al. (1997). "The AIMS2-SF: a short form of the Arthritis Impact Measurement Scales 2." Arthritis & Rheumatism **40**: 1267-1274.

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. (2003). A Comparison of Five Multi Attribute Utility Instruments, Working Paper 140. Melbourne, Centre for Health Program Evaluation, Monash University.

Helliwell, J., R. Layard, et al. (2012). World Happiness Report. New York, Earth Institute, Columbia University http://www.earthinstitute.columbia.edu/articles/view/2960 [Accessed 30 April 2012].

Höfer, S., L. Lim, et al. (2004). "The MacNew Heart Disease health-related quality of life instrument: A summary." Health and Quality of Life Outcomes **2**(3): 1-8.

Iezzi, A., D. Heriot, et al. (2012). Construction and validation of an online self TTO instrument for measuring health related quality of life (forthcoming). Melbourne, Centre for Health Economics, Monash University.

Kahneman, D., E. Diener, et al. (2003). Wellbeing: the Foundations of Hedonic Psychology. New York, Russell Sage Foundation.

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.

Kessler, R. C., P. R. Barker, et al. (2003). "Screening for serious mental illness in the general population." Arch Gen Psychiatry **60**(2): 184-189.

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 (2009). Report on Health Related Quality of Life and Lifestyle of Bangladeshi Migrants in Melbourne: Use of Multi-Attribute 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.

Laibson, D. and R. Zeckhauser (1998). "Amos Tversky and the ascent of behavioral economics." Journal of Risk and Uncertainty **16**: 7-47.

Layard, R. (2005). Happiness: Lessons from a New Science. London, Penguin Books.

Lohr, K., R. Brook, et al. (1986). "Use of medical care in the RAND Health Insurance Experiment. Diagnosis and service-specific anlayses in a randomized controlled trial." Medical Care **24**: S1-S87.

Lovibond, S. H. and P. F. Lovibond (1995). Manual for the Depression Anxiety Stress Scales. (2nd. Ed.). Sydney, Psychology Foundation.

Marks, G. B., S. M. Dunn, et al. (2006). "A scale for the measurement of quality of life in adults with asthma." Journal of Clinical Epidemiology **45**(5): 461-472.

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.

McDowell, I. (2006). Measuring Health: A Guide to Rating Scales and Questionnaires. Oxford, Oxford University Press.

McHorney, C., J. Ware Jr, et al. (1994). "The MOS 36 item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups." Medical Care **32**: 40-66.

Meenan, R. F., J. H. Mason, et al. (1992). "AIMS2: The content and properties of a revised and expanded Arthritis Impact Measurement Scales health status questionnaire." Arthritis & Rheumatism **35**: 1-10.

Meguro, M., E. A. Barley, et al. (2007). "Development and Validation of an Improved, COPD-Specific Version of the St. George Respiratory Questionnaire*." Chest **132**(2): 456-463.

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.

Nussbaum, M. (2000). Women and Human Development: The Capabilities Approach. New York, Cambridge University Press.

ONS (2012). Integrtaed Household Survey User Guide, Volume 1 IHS Background and Methodology, http://www.ons.gov.uk/ons/guide-method/method-quality/specific/social-and-welfare-methodology/integrated-household-survey/index.html [Accessed 22 May 2012].

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., J. McKie, et al. (2011). 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 [http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper64.pdf].

Robeyns, I. (2005). "Selecting capabilities for quality of life measurement." Social Indicators Research **74**: 191-215.

Sen, A. (1993). Capability and Well-Being. The Quality of Life. M. Nussbaum and A. Sen. Oxford, Clarendon Press.

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., 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.

Smith, D. M., S. L. Brown, et al. (2008). "Are subjective well-being measures any better than decision utility measures?" Health Economics, Policy and Law **3**: 85-91.

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.

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.