MONASH University
Business and Economics

Centre for Health Economics

Research Paper 2012 (77)

Why are we Weighting: Do utility weights improve the predictive power of multi attribute utility (MAU) instruments?

Professor Jeff Richardson

Foundation Director, Centre for Health Economics
Monash University

Dr Munir A Khan

Research Fellow, Centre for Health Economics
Monash University

September 2012
(Revised March 2014)

Centre for Health Economics
ISSN 1833-1173
ISBN 1 921187 76 X

Acknowledgments

This research was funded by National Health and Medical Research Council (NHMRC) Grant ID: 491162.

Correspondence:

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

Purpose

The aim of this paper is to investigate the importance of the utility formula of four multi attribute utility (MAU) instruments in explaining the different utilities which are predicted by them. Two hypotheses are tested. The first is that the use of the utility algorithms will result in greater convergence of scores than the use of unweighted instrument scores. The second and contrary hypotheses is that differences in the utilities predicted by different MAU instruments will be primarily explained by differences in the adjusted unweighted values implying an insignificant role for the effect of the utility algorithms.

Methods

Using results from a small pilot study, 'values' were obtained for four MAU instruments by assigning equal weights to the items in their descriptive systems. Values were subject to a simple linear transformation to align the scale with the scale of the utility instrument. Utilities and adjusted values were used to test the two study hypotheses.

Results

Utilities displayed no greater convergence than unweighted values. Differences in utilities were largely explained by differences in adjusted values.

Conclusion

Relatively greater effort should be given to the descriptive systems of MAU instruments. Greater sophistication in the derivation of utility formula is unlikely to significantly increase the correspondence of MAU instrument utilities.

TABLE OF CONTENTS

List of Tables

List of Figures

Why are we Weighting: Do utility weights improve the predictive power of multi attribute utility (MAU) instruments?

1 Introduction

Health related quality of life (HRQoL) can be measured by multi attribute (MA) instruments with either weighted or unweighted scores. Both have a 'descriptive system' or 'descriptive classification' which consists of a set of items – questions or statements concerning the quality of life (QoL), and a set of response categories. 'Unweighted' – multi attribute value (MAV) – instruments are generally favoured in the psychological literature. A score is obtained by assigning the same importance to each item (ie there are no variable weights) and summing the rank order of the responses to obtain an overall score or 'value' from the instrument. In contrast, the multi attribute utility (MAU) instruments used for the calculation of quality adjusted life years (QALYs) employ a utility algorithm to determine a unique weight for each health state. Weights seek to measure the strength of preference for a health state and, consequently, the utility weights convert health state descriptions in to health state utilities. The 'construct' which MAU instruments seek to measure therefore differs from 'health related quality of life' as it seeks to measure people's preferences whereas indices of HRQoL do not purport to do this. Even without this separate purpose.

There is a strong case for the use of importance weights. Health states consist of a number of dimensions (broadly, physical and psycho-social) and, if the number of items describing these is arbitrary, then the numbers produced by unweighted instruments will be arbitrary. If results from different MAV instruments are to be compared and interpreted as measuring the same construct (HRQoL) then there is a strong argument for weighting item responses to increase the relative importance of under-represented dimensions and to decrease the importance of dimensions with a relative abundance of items.

An additional reason for the use of utility weights is that across the spectrum of health states the relative importance of different items and dimensions can vary and a properly constructed set of weights can accommodate this. For example, impaired mobility may have a relatively large negative effect upon the utility of an otherwise healthy person. However, the importance of the same level of immobility may fall if the person is severely depressed and has no wish to be active. A flexible set of utility weights may take account of such interactions in a way which is not possible with equally unweighted items.

Despite these considerations, it is argued in the psychological literature that variable weights may not improve the performance of instruments. In a landmark article, Dawes (1979) argued that complex statistical algorithms add little to the predictive power of simple scoring methods, a view which has been subsequently defended theoretically and empirically (Trauer and Mackinnon

2001, Wu 2008). The theoretical arguments have drawn upon Locke's (1969, 1976) 'Range of Affect' hypothesis. This maintains that the response to satisfaction questions will reflect the importance of the subject to the individual even when there is no explicit reference to its importance in the question: people will take importance into account psychologically and give more extreme responses when the subject matter is of importance. Empirical evidence for the hypothesis has been found by Dana and Dawes (2004), Wu and Yao (2006, 2006) and Wu et al. (2009).

The subject matter of the psychologist's critique – satisfaction scales – differs in important respects from the subject matter of MAU scales. MAU scales make limited use of satisfaction questions and typically ask quasi-objective questions ('do you have problems walking; are you happy and interested in life?'). However, answers are not purely objective and response categories are often subjective ('a lot', 'a little', 'not at all'). In principle, this subjectivity makes responses vulnerable to the Range of Affects phenomena.

A second theoretical explanation for Dawes' conclusion is that utility weights derived from regression analysis may be non-optimal. Regression coefficients are unbiased but can be inefficient. Coefficients from a sub-sample of the total population may 'over-fit' the data by adjusting to best fit a specific sample. As a result there will be 'shrinkage' (a reduction in R^2) when results are applied to the full population or another sample (Guion 1965). For related reasons it has been argued that regression coefficients may not be the most efficient for achieving predictive validity (Gigerenzer and Todd 1999, Dana and Dawes 2004). Parameters obtained from any weighting methodology may not correctly represent the preferences of a subset of patients in a particular study. Summarising psychological research, Kahneman (2011) reports that 'formulas that assign equal weights to all the predictors are often superior because they are not affected by accidents of sampling' (p226). It is further suggested that for specific purposes – which, in the present context is the measurement of utility – a simple adjustment to the unweighted, global score may achieve equal or better results than the use of variable weights (Guion 1965).

In sum, there are theoretical reasons for the use of health state specific utility weights rather than a simple adjustment to the global score obtained from the use of equal weights. However there are also counter arguments and evidence for doubting the advantage of this approach. The purpose of the present article is to investigate this question using values and utilities from six MAU instruments. Data from these instruments were used to test the two study hypotheses below:

Hypothesis 1 Convergence: That the use of utility algorithms will result in greater convergence of instrument scores than the use of unweighted instrument scores.

Hypothesis 2 Prediction: That differences in the utilities predicted by different MAU instruments will be primarily explained by differences in unweighted data after a single adjustment to align measurement scales.

Four tests were conducted to determine the plausibility of these hypotheses.

Test 1 Convergence: Whether utilities derived from MAU instruments correlate more highly than the unweighted values derived from the same instruments: $\rho\ (U_i\ U_j)$ vs $\rho\ (V_i\ V_j)$.

Test 2 Convergence: Whether utilities derived from an MAU instrument correlate more highly with other instrument utilities than with the unweighted scores derived from these other instruments: $\rho\ (U_i\ U_j)$ vs $\rho(U_i\ V_j)$.

Test 3 Prediction: Whether variation in utilities derived from MAU instrument algorithms will be largely explained by variation in the unweighted values derived from the same instrument: (ρ (U_i V_i)).

Test 4 Prediction: Whether pairwise differences in utilities predicted by different MAU instrument algorithms will be largely explained by differences in the unweighted values after a simple adjustment to align overall measurement scales: $(U_i - U_j) = f(V_i - V_j)$.

Test 1: The reasons for the first hypotheses were discussed above. Instruments with differing dimensions and items per dimension would be expected to correlate poorly. The weights created for each of the MAU instruments are designed to convert the dimension descriptions into a single quantity, the health state utility. To the extent to which this has been achieved and weights compensate for differences between dimension structures, the weighted scores should correlate more highly than unweighted values.

Test 2: The logic of the second test is similar to the logic of discriminant validity. If utility weights successfully convert disparate scores into similar (and, ideally, identical) utilities, then the correlation between instrument utilities should be high and replacing one of the utilities with an unweighted value should reduce the correlation.

Test 3: Utility and unweighted values will correlate as each varies with the underlying health state. Nevertheless the lower the correlation between them, the greater the independent importance of the utility weights.

Test 4: Differences between utilities predicted by different instruments may be positive or negative and need not correlate with the corresponding difference between unweighted values. In the extreme, different instruments might produce identical values ($V_i - V_j = 0$). In this case differences in MAU scores ($U_i - U_j$) would be entirely a reflection of the importance of the utility weights. The fourth test therefore identifies the contribution of the health state specific utility weights to the discrepancies between predicted utilities. Good prediction of differences by unweighted values indicates that discrepancies primarily reflect differences in the descriptive system. Poor prediction implies the relative importance of differences in health state specific utility weights.

Values calculated from unweighted instruments are unlikely to fall on the same scale as utilities which are computed from techniques which seek to quantify the strength of people's preferences. The fourth test is therefore carried out with adjusted values derived from a linear transformation to the instrument value to align the value and utility scales. Comparisons are effectively between utilities estimated from flexible weights and simple utilities estimated from a single adjustment. However to distinguish these, the terminology 'adjusted value' is retained.

2 Methods

Data: Data were obtained from a pilot study which administered four MAU instruments – the EQ-5D, SF-6D, HUI 3 and AQoL-8D –to a group of relatively healthy residents from within the Bangladeshi community of Melbourne. An open invitation to participate in the project was distributed through community organisations, cultural groups, businesses and community leaders. Those expressing an interest were offered the option of a postal questionnaire and prepaid return envelope or a face-to-face interview at a location convenient to them. These included community venues or an individual's home. Upon receipt, questionnaires were checked. Data were then entered into SPSS for analysis. Details of the survey method and results are given in Khan and Richardson (2011). Data were collected with the approval of the Monash University Human Ethics Committee (CF08/28946 – 2008001494). The tenets of the Declaration of Helsinki were adhered to.

The four instruments are summarised in Table 1 and described in Brazier al. (2007) and Richardson et al. (2011). Instruments differ in their conceptualisation, size, content and scoring formula. While the HUI 3 has a 'within the skin' descriptive system (which focuses upon an individual's body functions) the other three instruments are based primarily, but not exclusively, upon handicap (more recently described by the WHO (2001) as 'activity' and 'participation') – a description of the effect of a health state upon a person's ability to function in a social environment. The SF-6D and AQoL-8D employed psychometric methods in the derivation of their items. HUI 3 and EQ-5D employed judgement and importance ranking. The items combine to describe between 243 health states (EQ-5D) and 8.7 x 10^{23} health states (AQoL-8D). The techniques used to derive utilities also differed. EQ-5D and SF-6D employed econometric methods. HUI 3 and AQoL-8D used the multiplicative formula recommended by decision analytic theory. The AQoL-8D applied a second stage econometric correction.

Table 1 Summary of 4 MAU instruments: Items per dimension [(a)]

Dimension	EQ-5D[(b)]	HUI 3	SF-6D (36)	AQoL-8D
Physical				
Physical Ability/Mobility	*	**	*	**
Vitality			*	*
Bodily Function/ Self Care	*			*
Pain/Discomfort	*	*	*	**
Senses/communication		***		***
Usual activities/role	*		*	****
Vitality				*
Psycho-Social				
Sleeping				*
Depression/Anxiety/Anger	*	*	*	******
Cognition/Memory/Ability		*		
Satisfaction/Happiness				****
Self Esteem				**
Social Function/Family			*	********
Summary: Total number of items	5 items	8 items	6 items	35 items

Notes:
*[(a)] *=1 item (ie question, response)*
[(b)] EQ-5D 3 level

Source: Brazier et al. (2007); Richardson et al.(2011).

A two stage method was used to calculate values, V, from the unweighted MAV instruments. In stage 1, item scores were set equal to the rank order of the response and then summed to obtain a score, X. Next, X was constrained to the range (0-1) using equation 1 below, where X_{min} and X_{max} are the scores obtained when the response to every item of the instrument is at its minimum (best) and maximum (worst) level respectively. Equation 1 produces a 'disvalue' score, DV – high scores correspond with poorer health. This was converted into a value score, V*, using equation 2. With the 35 item AQoL-8D this procedure was carried out for each of the 8 dimensions and the 8 dimension scores were averaged.

$$DV = (X-X_{min})/(X_{max}-X_{min}) \quad \ldots \quad \text{equation 1}$$

$$V^* = 1-DV \quad \ldots \quad \text{equation 2}$$

As an example of this first stage, if the responses to the (3 level) EQ-5D were (1, 3, 3, 2, 2), then the sum of the rank scores would be 1+3+3+2+2=11. The maximum and minimum Stage 1 scores for the EQ-5D-3L were therefore 15 (3 x 5) and 5 (1x5) respectively. Consequently from equation 1 the disvalue DV would be (11-5)/(15-5)=0.6 and from equation 2 the value of the health state V* would be 1-0.6=0.4.

In the second stage values V* were subject to a simple linear transformation to align the scale with the scale used by the corresponding utilities. Parameters for the transformation were obtained by regressing each instrument's utility scores, U, on the corresponding value V* as shown in equation 3.

$$U = a + bV^* + e \quad \ldots \quad \text{equation 3}$$

The parameters a and b were then used to replace V* with V using equation 4.

$$V = a + bV^* \quad \ldots \quad \text{equation 4}$$

Substituting V* from equation 4 in equation 3:

$$U=a+b\,(V-a)/b$$

$$U=0.0+1.00V$$

Figure 1 illustrates the second stage. If equation 1 is U=0.62 + 0.38 V (which is reported later for HUI 3) then the transformation equation 4 is V=0.62 + 0.38V or V*=1.63 + 2.63 V. This rotates the relationship between V* and U so it corresponds with the line U=V.

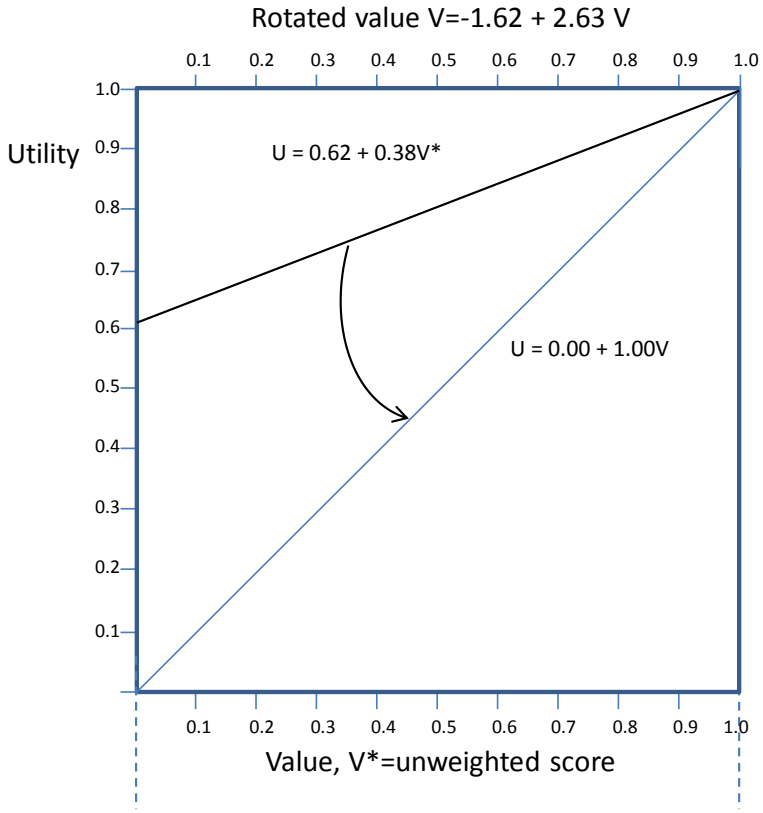
A problem with this rotation is that equation 3 may not be estimated when utility scores are unavailable. This implies that this method could not be used to adjust values and replace utilities. However the transformation function does not require a full set of utility weights and can be estimated with reliable information on the value of a single point. In the small sample available for this study no such reliable value is available (ie for a single point). Consequently the alternative is illustrated using the average utility and value scores for the lowest 40 observations for each instrument.

Convergence: Convergence was tested using Pearson correlation. The first test was a comparison of the correlation between utilities and between values. The study hypothesis is that the former correlations will not exceed the latter. The second test was a comparison of the correlation between utilities with the correlation between utility and value. The study hypothesis was, again, that the former correlation would not exceed the latter.

Prediction: Prediction was tested using Geometric Mean Squares (GMS) regression. These obtain their parameters from the geometric mean of the parameters of OLS regression of X on Y

and Y on X. Consequently they allow for error terms in both variables and results are independent of the choice of 'dependent' and 'independent' variables (Tofallis 2002). The third test was to determine the explained, relative to the unexplained, variance in the regression of utility upon value from the same instrument. The fourth test was to compare explained and unexplained variance in the regression of pairwise differences in utilities upon differences in values.

Figure 1 Example of a linear transformation



3 Results

A total of 158 individuals completed the questionnaire. Responses from five individuals had internal inconsistencies and were deleted. Table 2 indicates that the overall age-gender composition was similar to the Australian profile. However, in other respects the respondents were atypical. All were first or second generation Bangladeshi migrants, 91 percent had a tertiary qualification, 85 percent lived in families including parents and/or spouse and/or children, and 89 percent reported being in good, very good or excellent health.

Summary statistics for the instruments are shown in Table 3. The high mean utilities for every instrument are indicative of the good health of the sample population. Mean utilities are similar for the EQ-5D, HUI 3 and lower for the SF-6D and the AQoL-8D, reflecting their inclusion of a larger proportion of psycho social items which are more likely to affect a relatively healthy population. However the similarity in the means conceals significant differences in the distribution of utility scores. The EQ-5D and HUI 3 have significant ceiling effects with 57.6 percent of respondents obtaining a score of 1.00. In contrast only 20.3 and 1.3 percent of SF-6D and AQoL-8D scores were equal to 1.00 respectively. The range of utilities varied from 0.32 for the EQ-5D to 0.57 for the AQoL-8D. Mean values are also similar except for the AQoL-8D, whose mean reflects the preponderance of psycho-social items. The range of instrument values varies from 0.22 for the HUI 3 to 0.64 for the EQ-5D.

Table 2 Age gender composition of the survey participants

Age group	Percent		Number	Percent 18-64	
	Male	Female	Total	% Sample	Australia
18-24	15.3	13.7	23	14.6	11.3
25-34	37.6	32.9	56	35.4	22.3
35-44	17.6	26	34	21.5	24.6
45-54	27.1	26	42	26.6	23.5
55-64	2.4	1.4	3	1.9	18.3
Total	100	100	158	100	100.0
n	85	73			

Table 3 Summary Statistics from the 4 MAU instruments (n=153)

MAUI	Metric	Mean	SE	Min	Max
EQ-5D	Utility (U)	0.92	0.008	0.68	1.00
	Value (V)	0.89	0.012	0.36	1.00
HUI 3	Utility (U)	0.91	0.009	0.51	1.00
	Value (V)	0.95	0.005	0.78	1.00
SF-6D	Utility (U)	0.87	0.008	0.60	1.00
	Value (V)	0.91	0.006	0.69	1.00
AQoL-8D	Utility (U)	0.85	0.009	0.43	1.00
	Value (V)	0.78	0.008	0.47	1.00

Notes:

U = Utility; V = Value

Convergence: The Pearson correlations needed for the first two tests are reported in Table 4. Correlations between instrument utilities are shown in the top left block of the results; between instrument values in the bottom left hand block and between utilities and values in the right hand block. Comparing the correlation of utilities with the correlation of values (Test 1) contradicts the first hypothesis. Each of the correlation coefficients between utilities in the top block is less than the corresponding correlation between values in the bottom left block with the exception of the correlations between AQoL-8D and both the EQ-5D and HUI 3 where the differences were statistically insignificant.

Comparing the correlations between utilities with the correlations between values (Test 2) similarly contradicts the hypothesis. The average correlation between the EQ-5D utility and values from other instruments – 0.57 – exceeds its average correlation with other utilities – 0.56. Similarly the average correlation of SF-6D utility with values – 0.62 – exceeds it correlation with other utilities –0.58. The average HUI 3 utility-value correlation – 0.56 – exceeds the average HUI 3 utility-utility correlation of 0.53. AQoL-8D is again the exception with the average correlation between utilities and values – 0.56 being marginally but insignificantly below the average utility-utility correlation of 0.58.

Table 4 Hypothesis 1A, 1B Pearson correlation between instruments' utilities and values

	Top left: Correlation between Utilities					**Correlation between Utilities, Values**				
	EQ-5D	SF-6D	HUI 3	AQoL-8D	Ave.	EQ-5D	SF-6D	HUI 3	AQoL-8D	
EQ-5D		0.57	0.50	0.61	0.56		0.57	0.59	0.56	**UTILITIES**
SF-6D	0.60		0.56	0.59	0.58	0.56		0.58	0.61	
HUI 3	0.60	0.61		0.53	0.53	0.52	0.64		0.51	
AQoL-8D	0.56	0.66	0.49		0.58	0.63	0.66	0.52		
Average	0.59	0.62	0.57	0.57		0.57	0.62	0.56	0.56	
	Bottom Left: Correlation between Values					**VALUES**				

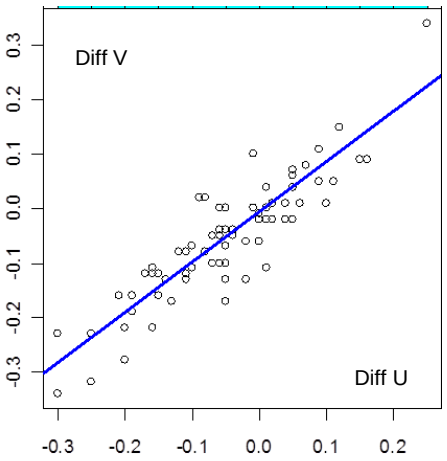
Prediction: Table 5 columns (a) reports the results of the GMS regression of utilities upon the stage 1 values, V*, derived from equation 3. They simultaneously provide the linear transformations needed to estimate adjusted values, V, in equation 4 and indicate the percentage of the variance in utilities which may be explained by values. (The per cent is unaffected by the use of V or V* as one is a linear function of the other.) Values explained between 88 and 92 percent of variance in utility. Between 8 and 12 percent of variance may therefore be attributed to the use of health state specific utilities and to random error.

Table 5 columns b report the linear transformation which would be obtained by algebraically estimating a line which passed through (0.00), (1.00) and a single point which was obtained for each instrument by averaging the lowest 40 observations and points are reported in Footnote 2. For three of the four instruments the simple approximation is virtually identical to the regression result.

Adjusted values, V, were employed in the GMS regressions reported in Table 6 and depicted in Figure 2. Results indicate that in each pairwise comparison differences in values explain between 81 and 90 percent of the variance in the difference between utilities, leaving 10-19 percent of variance to be explained by health state specific utility weights and random error. Slope coefficients were within 10 percent of unity: a given difference in adjusted values results in an almost identical difference in utilities.

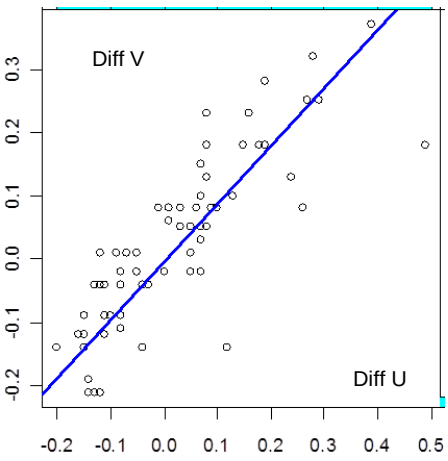
Figure 2 Difference in utilities upon difference in values

SF-6D-EQ-5D



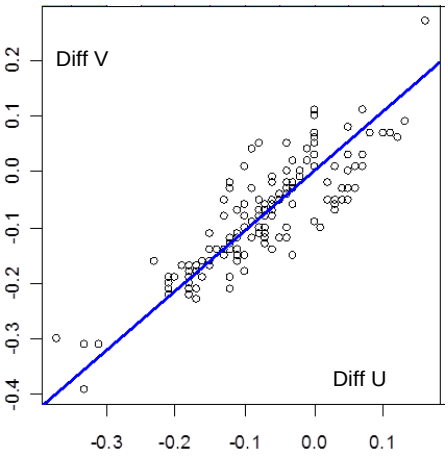
Diff U = -0.005 + 0.926 Diff V R^2=0.9

EQ-5D-HUI 3



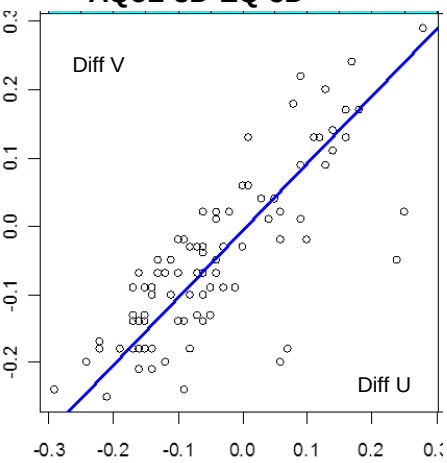
Diff U = -0.004 + 0.913 Diff V R^2=0.83

SF-6D-HUI 3



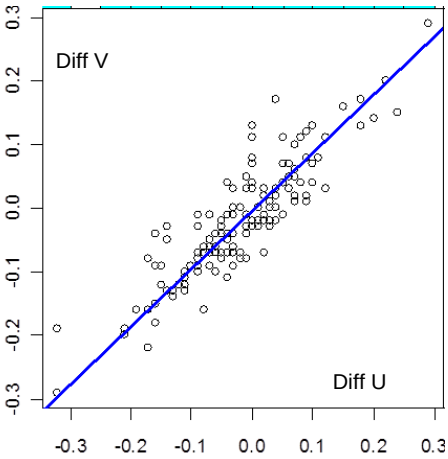
Diff U = -0.006 + 0.988 Diff V R^2=0.81

AQoL-8D-EQ-5D



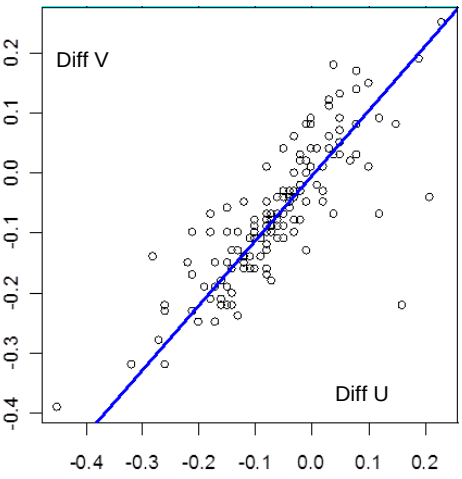
Diff U = -0.001 + 1.07 Diff V R^2=0.86

AQoL-8D-SF6D



Diff U = -0.004 + 0.915 Diff V R^2=0.88

AQoL-8D-HUI 3



Diff U = -0.004 + 1.08 Diff V R^2=0.82

Table 5 Regression of utility, U, on Value, V*[(a)]

	EQ-5D[(b)]		SF-6D[(b)]		HUI[(b)]		AQoL-8D[(b)]	
	A	B	A	B	A	B	A	B
Constant	0.04	0.05	0.07	0.07	-1.62	-1.62	-0.48	-0.27
b	0.95	0.95	0.91	0.93	2.63	2.63	1.60	1.27
R^2	0.88		0.92		0.88		0.92	

Notes

(a) The independent variable V is the first unweighted value described in the text defined by equations (1) and (2).

(b) Column (A) reports the results from (GMS) regression of the instrument's utilities, U, upon value, V Column (B) reports the equations fitted through (0.0), (1.00) and a single point which was obtained for each instrument by averaging the lowest 40 observations. The (U,V) scores for these points were: EQ-5D (0.79, 0.78); SF-6D (0.72, 0.70); HUI 3 (0.66, 0.87); AQoL-8D (0.60, 0.68).*

Table 6 Regression of (U_i-U_j) upon (V_i-V_j)*

(U_i-U_j) on (V_i-V_j)	a	b	R^2	(U_i-U_j) on (V_i-V_j)	a	b	R^2
i=SF-6D:j=EQ-5D	0.0	0.93	0.9	i=EQ-5D:j=HUI 3	0.0	0.91	0.83
i=SF-6D:j= HUI 3	0.00	0.99	0.81	i=AQoL-8D:j=HUI 3	0.00	1.07	0.86
i=AQoL-8D:j=SF-6D	0.00	0.92	0.88	i=AQoL-8D:j=HUI 3	0.00	1.08	0.82

** Geometric mean squares regression*

4 Discussion

The quality adjusted life year was introduced into economic evaluation to allow outcomes where the QoL differed to be measured in comparable units. The utility weights employed to create QALYs seek to measure the strength of preference for different health states.

Multi attribute utility instruments were introduced to facilitate the measurement of these utilities. However the comparability of results from different instruments remains a problem. Drummond for example, notes that 'These systems are far from identical. They differ in the dimensions of health covered, in the number of levels defined on each dimension, in the description of these levels… Because of these various differences it is not surprising that comparative studies show that the same patient groups can score quite differently depending upon the instrument used'. (Drummond, Sculpher et al. 2005 pp160-170). Similarly Brazier et al.(2007) argue that 'generic measures of health have been found to be inappropriate or insensitive for many medical conditions… generic measures are designed to cover the core dimensions of health… no instrument is able to cover all health dimensions (pp60-63).

In principle, these differences may be mitigated through the use of flexible health state specific utility weights which increase the importance of under-represented elements of dimensions and reduce the importance of others. The very significant research effort devoted to the creation of improved and country specific utility weights appears to reflect a belief that precision in utility formulae is a priority task for the prediction of valid utilities from MA instruments.

The present paper has tested this expectation against the contrary view found in the psychological literature that weights will not necessarily achieve greater validity and, by extension, that the focus of research should be upon instrument descriptive systems. Instrument values derived from unweighted instrument items and unrelated to patient preferences cannot represent utility. However the research question here has been whether or not a simple linear transformation of the value scale can produce adjusted values which represent utility as well as or better than existing utility algorithms. The results suggest that this may be true.

The results are tentative and are limited for two major reasons. The first is that the sample population was small, atypical and relatively healthy. There are no strong reasons for believing that the relationship between weighted and unweighted instruments should vary with ethnicity and education. However it might vary significantly with the health and type of ill health of respondents. For this reason the study needs replication with a larger sample.

The second caveat is that the adjustment to the values obtained from unweighted instruments was based upon information obtained from the utility instruments. This does not detract from the test results for the first hypothesis relating to convergence. The correlations reported do not vary with a linear transformation. The adjustment carried out to test the second hypothesis may, in principle, be based upon a single point or a limited number of observations. The use of a single point for this purpose, reported in Table 5, resulted in linear transformations between utility and unadjusted values, V*, which were effectively identical to the regression results for three of the four instruments.

5 Conclusion

Despite the caveats above, this paper raises questions which have not been discussed in the health economics literature. The *prima facie* case for using health state specific utility weights is so strong that the case presented in the psychology literature has been ignored. However results here support the psychologists' contention that the health state specific weighting of instrument values may not improve their convergent or predictive validity.

If the present findings are confirmed in subsequent studies they have important implications for future research. At present the major focus in the literature is upon the improvement of utility weights. The present paper adds to the evidence that this is not the most pressing issue and that the chief focus should be upon instrument descriptive systems. More controversially, it suggests that greater reliability and validity might be achieved with a simple adjustment to unweighted scales than with the use of increasingly sophisticated techniques for the creation of utility formula.

References

Brazier, J., J. Ratcliffe, J. Salomon and A. Tsuchiya (2007). Measuring and Valuing Health Benefits for Economic Evaluation. Oxford, Oxford University Press.

Dana, J. and R. Dawes (2004). "The superiority of simple alternatives to regression for social science predictions." Journal of Educational and Behavioral Statistics **29**(3): 317-331.

Dawes, R. (1979). "The robust beauty of improper linear models in decision making." American Psychologist **34**(7): 571-582.

Drummond, M., M. Sculpher, G. Torrance, B. O'Brien and G. Stoddart (2005). Methods for the Economic Evaluation of Health Care Programs. Oxford, Oxford University Press.

Gigerenzer, G. and P. M. Todd (1999). Simple Heuristics that Make us Smart. London, Oxford University Press.

Guion, R. M. (1965). Personnel Testing. New York, McGraw-Hill.

Kahneman, D. (2011). Thinking Fast and Slow. New York, Farrar Straus & Giroux.

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.

Locke, E. (1969). "What is job satisfaction?" Organizational Behavior and Human Performance **4**: 309-336.

Locke, E. (1976). The nature and causes of job satisfaction. Handbook of Industrial and Organizational Psychology. M. D. Dunnett. Chicago, Rand McNally.

Richardson, J., J. McKie and E. Bariola (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.

Tofallis, C. (2002). Model Fitting for Multiple Variables by Minimising the Geometric Mean Deviation. Total Least Squares and Errors-In-Variables Modeling: Algorithms, Analysis and Applications. S. Van Huffel and P. Lemmerling, Kluwer Academic. Available at SSRN: http://ssrn.com/abstract=1077322.

Trauer, T. and A. Mackinnon (2001). "Why are we weighting? The role of importance ratings in a quality of life measurement." Quality of Life Research **10**: 579-585.

WHO (2001). International classification of functioning, disability and health. Geneva, World Health Organization http://apps.who.int/classifications/icfbrowser/ [Accessed 27 July 2013].

Wu, C. (2008). "Examining the appropriateness of importance weighting in satisfaction score from range-of-affect hypothesis: hierarchical linear modeling for within-subject data." Social Indicators Research **86**: 101-111.

Wu, C., L. H. Chen and Y. Tsai (2009). "Investigating importance weighting of satisfaction scores from a formative model with partial least squares analysis." Social Indicators Research **90**: 351-363.

Wu, C. and G. Yao (2006). "Do we need weight item satisfaction by item importance? A perspective from Lock's Range-of-Affect hypothesis." Social Indicators Research **79**: 485-502.

Wu, C. and G. Yao (2006). "Importance has been considered in satisfaction evaluation: an experimental examination of Locke's Range-of-Affect Hypothesis." Social Indicators Research **81**: 521-541.