MONASH University
Business and Economics

Centre for Health Economics

Working Paper 03/14

Do Utility Formulae Accentuate or Diminish Differences between Multi Attribute Utility (MAU) Instruments

Professor Jeff Richardson

Foundation Director, Centre for Health Economics
Monash University

Angelo Iezzi

Research Fellow, Centre for Health Economics
Monash University

October 2014

Centre for Health Economics
ISSN 2204-0218
ISBN 1 921187 89 1

Do utility formulae accentuate or diminish difference between multi attribute utility (MAU) instruments

Jeff Richardson*[1], Angelo Iezzi*

*Centre for Health Economics, Monash University

30 September 2014

Abstract

Multi attribute utility (MAU) instruments generate an index of utility by summing the weighted responses to a health related quality of life questionnaire or by applying a formula to the responses. In principle, this converts dissimilar attributes – the level of pain, mobility, depression, etc – to a homogeneous unit, utility; ie the strength of preference for the health state described. Adding up scores from different attributes without weighting them appears to be like adding apples and oranges. In contrast, it has been argued by psychologists that the methods used to construct weights are likely to introduce bias and more reliable results may be obtained from 'unweighted' instruments, ie instruments which apply the same weight to every response. In effect, the argument is that the benefit of using equal, untampered weights is greater than the benefit of using weights derived with problematical methods.

The present paper tests these conflicting views. Pairwise differences between the utilities predicted by five MAU instruments are decomposed into three explanatory effects: differences in their measurement scales, their descriptive systems and the residual 'micro utility effect'. Descriptive systems differ significantly. The question investigated here is whether or not the utility formulae help to convert these different descriptive systems into a homogeneous metric, viz, utility: that is, do the micro utility effects of the formula reduce differences between the values which would be obtained by adding up scale adjusted but unweighted scores. In the event, it is found that micro utility effects are small but positive, supporting the psychologists' hypothesis that, after adjusting for scale, utility formula accentuate rather than diminish differences between the utilities predicted by MAU instruments.

[1] Corresponding author: Jeff Richardson; jeffrey.richardson@monash.edu; +61 399050754; Centre for Health Economics, Building 75, Monash University, Clayton 3800, Victoria, Australia.

Do utility formulae accentuate or diminish difference between multi attribute utility (MAU) instruments

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 (MAU) – 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.

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 & 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 (2006a, b) and Wu et al. (2009).

The second supporting argument for Dawes' conclusion is that utility weights derived from regression analysis may 'over-fit' the data by adjusting to 'best fit' a specific sample (Guion, 1965). For related reasons it has been argued that regression coefficients may not be the most efficient for achieving predictive validity (Dana & Dawes, 2004; Gigerenzer & Todd, 1999). 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 argues (2011) that 'formulas that assign equal weights to all the predictors are often superior because they are not affected by accidents of sampling' (p226). Kahneman further suggests 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. Following this suggestion, in the present study a simple adjustment is made to the unweighted scores to ensure that adjusted scores and utilities are on the same linear scale. Such unweighted but transformed scores will be referred to here as 'values'.

The differences between MAUI utilities are necessarily the result of the differences between the MA descriptive systems and the instruments' utility formulae. By assigning numbers to the health state descriptions, the utility formulae determine the measurement scale. The focus of this paper is to determine the extent to which these formulae, additionally, contribute towards or diminish the differences between the utilities predicted by MAUI; that is to differences in utilities which are less than the differences between scale adjusted values derived from unweighted scores. A diminution in the difference would be evidence of convergent validity; that the instruments were measuring the same quantity. If the formulae

increase, not diminish differences, then the result supports the psychologist's contention that weighting detracts from instrument validity.

Methods

Pairwise differences in utilities were decomposed into their three component parts, ie scale, descriptive system and micro utility components as shown in Figure 1. This plots scores, S_i, S_j, derived by summing item responses from two MAUI, $MAUI_i$ and $MAUI_j$ on the horizontal axis, and the corresponding utilities, U, and values, V, on the vertical axis. Values are a linear transformation of scores and are represented by the lines XY, ZY. Due to the micro utility effects of the MAU formula the corresponding instrument utilities are scattered randomly around the two lines. The differing measurement scales embodied in the utility formula are illustrated by the differing slopes of XY and ZY. For a given individual, A, the scores from the unweighted instruments S_i^A, S_j^A will differ. Application of the two MAUI formulae results in estimates of utility which differ by $(U_i^A-U_j^A)$. The aim of the analysis below is to attribute this difference to a difference in the scale $(V_i^A - V_j^A)$, a difference in the micro utility effect $(V_i^A-U_i^A)$ and $(V_j^A-U_j^A)$ and the effect attributable to the structure of the descriptive systems which results in the difference, $S_i^A-S_j^A$.

Analysis: Terminology is defined in Box 1. For each respondent absolute (sign free) differences (U_i-U_j) were calculated for each instrument pair. (Consequently, two differences of -0.6 and +0.4 will average 0.5, not 0.1.) A two stage method was used to calculate values, V_i. In stage 1, the rank order of item responses were summed to obtain an initial 'rank order' score, R. This was constrained to the range (0-1) to obtain a score, S, using equation (1).

$$S_i = (R_i-R_{min})/(R_{max}-R_{min}) \qquad \ldots \text{ equation 1}$$

where R_{min}, R_{max} are the minimum and maximum 'rank order' scores which may be obtained from the instrument.

In the second stage, scores, Si, were subject to a linear transformation to obtain 'values' which were calibrated on the same scale as the corresponding utilities (XY, ZY in Figure 2). An OLS linear regression, equation 2, was estimated for each instrument between utilities, U_i and scores S_i

$$U_i = a+b\ S_i + res_i \qquad \ldots \text{ equation 2}$$

Values, V, were calculated by deleting the residual, res_i, ie $V_i = a + b\ S_i$. Values calculated in this way are therefore a linear transformation of unweighted scores. Utilities, U_i, determine the scale upon which values V_i are calibrated. Values differ from utilities by the 'micro utility effect' included in res_i.

In each pairwise comparison of MAU_i and MAU_j the effect of scale was removed by similarly rotating U_j and V_j to be on the same scale as U_i. This was achieved by regressing both U_j and V_j upon U_i as shown in equations 3 and 4.

$$U_i = a_1+b_1\ U_j + res_1 \qquad \ldots \text{ equation 3}$$

$$U_i = a_2+b_2\ Vj + res_2 \qquad \ldots \text{ equation 4}$$

where res_1, and res_2 are residuals attributable to micro utility effects and measurement error.

Rotated utilities and values were obtained from the linear component of these equations as defined by equation 3' and 4'.

$$U_j^* = a_1+b_1\ U_j \qquad \ldots \text{ equation 3'}$$

$$V_j^* = a_2+b_2\ V_j \qquad \ldots \text{ equation 4'}$$

where U_j^* and V_j^* are respectively the utility and value from MAU_j rotated to be on the same scale as U_i. The effect of the linear adjustment (3') may be shown by substituting $U_j=[U_j^*-a_1]/b_1$ derived from equation 3' into equation 3.

$$U_i = a_1+b_1[U_j^* - a_1]/b_1 + res_1$$

$$U_j^* = U_i - res_1 \qquad \ldots \text{ equation 5}$$

Similarly, substituting $V_j = [V^*-a_2]/b_2$ from equation 4' into equation 4

$$U_i = a_2+b_2\ [V_j^*-a_2]/b_2 + res$$

$$V_j^*= U_i - res_2 \qquad \ldots \text{ equation 5'}$$

Equation 5 and 5' confirm that U_j^* and V_j^* are on the same linear scale as U_i, varying from U_i by res_1 and res_2 respectively, which include the effects of differing descriptive systems micro utility effects and an error term.

To test the success with which scale effects were removed by these procedures, OLS regressions were estimated between differences in the scale adjusted utilities and values: equation 6. With linear relationships between variables a perfect alignment of scales would result in $a_3 = 0$; $b_3 = 1.00$. Non-linearities in the relationships would result in $a_3 = 0$ (a property of OLS regression) but possible deviation from $b_3 = 1.00$.

$$[U_i - U_j^*] = a_3 + b_3 [V_i^* - V_j^*] \quad \text{… equation 6}$$

Disaggregation employed the following relationships:

$A = U_i - U_j$: Pairwise difference in utilities which are to be explained

$B = U_i - U_j^*$: *'Scale free' differences in utility.* The differences in utility measured on a common scale (MAU_i).

$C = A - B$: *The scale effect.* The amount of the difference, A, explained by measuring differences on a common scale.

$D = V_i - V_j^*$: *Descriptive system effects.* The scale free difference in values attributable (only) to differences in the descriptive system.

$E = B - D$: *The micro utility effect.* The scale free differences in utility less the scale free differences in V.

Combining the effects:

Scale (C) + Descriptive system (D) + micro utility (B-D)

$= C + D + B - D$

$= C + B = (A - B) + B = A = U_i - U_j$

Data: A Multi Instrument Comparison (MIC) survey was carried out in six countries: Australia, Canada, Germany, Norway, the UK and the USA. The online survey was administered by a global panel company, CINT Pty Ltd. The survey was approved by the

Monash University Human Research Ethics Committee (MUHREC), Monash University Melbourne Australia, reference number CF11/3192-2011001748.

Respondents were initially asked to indicate if they had a chronic disease and to rate their overall health on a VAS where 0.00 represented death and 100 represented 'best possible health' (physical, mental and social)'. Quotas were then used to obtain a demographically representative sample of the 'healthy' public, defined by the absence of chronic disease and by a score above 70 on the numerical health scale. Quotas were also applied to obtain a target number of respondents in each of seven chronic disease areas, viz, arthritis, asthma, cancer, depression, diabetes, hearing loss and heart disease.

The MAUI included in the study are described in Table 1. For four instruments, utilities were calculated using algorithms provided by the instruments' authors: SF-6D (Brazier et al., 2002), HUI 3 (Feeny et al., 2002), 15D (Sintonen & Pekurinen, 1993), and AQoL-8D (Richardson et al., 2014b). The 5 level EQ-5D-5L utilities were obtained from the crosswalk published by the EuroQoL Group (Rabin et al., 2011), derived using methods described by van Hout et al. (2012).

Responses were subject to a set of stringent edit procedures based upon a comparison of duplicated or similar questions. Additionally, results were removed when an individual's (recorded) completion time fell below 20 minutes which was judged to be the minimum time in which the 230 questions could be answered. Edit procedures, the questionnaire and its administration are described in Richardson et al. (2012).

Results

Data: Data were obtained from 9,665 individuals. Edit procedures resulted in the removal of 17 percent of the total. Table 2 presents the age-gender and educational status of the remaining 8,022 respondents. Because quotas were imposed the proportion of respondents from each country is similar. For the same reason, the age, gender and educational profiles of respondents within each country is similar. The numbers recruited from the disease area varied from 772 for cancer to 943 for heart disease. The 1760 'public' respondents were obtained by combing country samples which closely matched the age-gender profile in each country. There were few missing data as the online program did not permit respondents to

proceed until questions were completed. Individuals who did not answer the final question were excluded. This resulted in a final sample of 8,022. Details of the sample administration and editing in each country are provided in country specific reports (Richardson et al., 2012b-g), and a detailed comparison of instrument differences in Richardson et al. (2014a).

Table 3 reports summary statistics for the five instruments and the correlation between utilities and values. With the exception of the 15D mean utilities are similar, varying from 0.83 to 0.88 in the public sample and from 0.68 to 0.85 in the full sample. Despite this similarity, the distribution of utilities differs significantly. In the full sample the standard deviation of the observations varies by 100 percent from 0.27 for HUI 3 to 0.13 for 15D and 0.14 for SF-6D. Ceiling effects (U = 1.00) vary from 19.1 percent (EQ-5D) to 0.3 percent (AQoL-8D) and the percentage with a utility below 0.4 varies from 0.3 for the 15D and 1.3 percent for the SF-6D to 13.9 percent for HUI 3 and 14.7 percent for AQoL-8D. Values obtained from unweighted scores necessarily have the same means as utilities as they were obtained from the regression of utilities upon scores. However, as utilities are not a linear function of scores, the range of values differs from the range of utilities. Nevertheless the correlation between values and utilities is very high, exceeding 0.89 in all cases and rising to 0.99 for the 15D.

Rescaling: The linear regressions used to rotate the scales of utilities and values are reported in Table 4. The 'b' coefficient indicates the extent to which, on average, incremental change in the 'independent' (RHS) instrument utility or value must be compressed or expanded to be on the same scale as the 'dependent' (LHS) instrument. From the regression between HUI 3 and 15D utilities, increments of the 15D utility must be expanded by a factor of 1.75 for equivalence with the HUI 3 scale. In contrast, increments of utility on the AQoL-8D must be compressed by a factor of 0.47 for equivalence with incremental utilities measured by the 15D.

The test of the success of the rescaling of instruments is reported in Table 5. Reflecting the properties of the OLS regressions used to rotate the scales, a = 0 in every regression indicating that each of the variables used in the regressions has the same mean (equal to the mean of U_i). In each case the slope parameter, b, is close to but deviates from 1.00 reflecting non-linearities in the relationship. In the decomposition of effects, the imperfect alignment of scales will result in an increased micro utility effect.

Decomposition: The decomposition of the pairwise differences in utilities is reported in Table 6. The average absolute difference between pairs of instrument utilities is 0.135. It varies from 0.114 (SF-6D, AQoL-8D) to 0.175 (15D, AQoL-8D). The largest component is the effect of the descriptive system which accounts for 66.0 percent of the difference; varying from 27.4 percent (15D, AQoL-8D) to 101.6 percent (HUI 3, AQoL-8D). Scale affects average 30.3 percent of the difference varying from 3.5 percent (EQ-5D, SF-6D) to 69.7 percent (15D, AQoL-8D). Micro utility effects are the smallest component, averaging 3.7 percent of the difference and the absolute value varying from 0.8 percent (EQ-5D, HUI 3) to 19.8 percent (EQ-5D, SF-6D).

Discussion

Different MAUI predict different utilities for three reasons. First, their measurement scales differ. Second, the structure and content of their descriptive systems differ and, third, utility formulae introduce differences which are not a result of scale effects.

The chief conclusion from the disaggregation presented here is that differences in utilities are primarily the result of differences in the descriptive systems. The conclusion is, arguably, unsurprising. However, this was not a necessary conclusion. Health related QoL may be conceptualised and described in different ways. Instruments which differ superficially may, in principle, give similar answers. Nevertheless this does not appear to occur with MAU instruments.

While descriptive systems explain 66.0 percent of the difference between utilities (U_i-U_j), their importance in pairwise comparisons varies from 27.4 percent in the comparison of the 15D and AQoL-8D to 101.6 percent of the difference between HUI 3 and AQoL-8D. The former results are plausible. As scale effects account for a larger part of the difference between 15D and AQoL-8D than for any other instrument pair, the relative importance of the remaining effects is consequently reduced. In Table 1 the 15D descriptive system uniquely shares with AQoL-8D items relating to sleep and intimacy and the two instruments have the largest number of items describing depression and anxiety. In contrast, the 'within the skin' descriptive system of HUI 3 has no items relating to social relationships which constitute a major part of the AQoL-8D descriptive system.

The more surprising result is that the principle effect of differing utility weights is via their effect upon measurement scales and not upon the micro utility effect. The scale effects are large in comparisons involving 15D and, from Table 3, the 15D has the lowest standard deviation implying the greatest compression of utilities. Scale effects are also large in the comparison of SF-6D with both HUI 3 and AQoL-8D. From Table 3, the SF-6D has the second lowest standard deviation and the HUI 3 and AQoL-8D have the largest standard deviations.

After taking account of differences in the descriptive system and scale, the residual micro utility effect – the motivating issue for the present paper– is generally positive: the effect contributes to, rather than diminishes, differences. In three cases in Table 6 it is negative suggesting that the effect partially compensates for other differences. With one exception the effect is small. The exception is the estimated micro utility effects in the comparison of EQ-5D and SF-6D. From Table 3 the relationship between SF-6D and EQ-5D is particularly non-linear with a rapid decrease in SF-6D utilities at the top end of the scale where 19 percent of EQ-5D utilities but only 1.3 percent of SF-6D are equal to 1.00. The pattern reverses as health deteriorates with 1.3 and 8.9 percent of observations below 0.4 for the SF-6D and EQ-5D respectively. Using present methods, the effect of non-linearities in the relationship between utilities is attributed to the micro utility effect.

Two related questions were posed in the introduction. The first was the extent to which, after adjustment for scale, utility formula added to or diminished the differences between the utilities predicted by MAUI. The results suggest that, on average, the utility formula accentuate rather than diminish differences, ie the micro utility effect is positive. This is consistent with the psychologists' hypothesis that results derived from different datasets and possibly over-fitted to particular models may reduce, not increase, the validity of measurement.

The second question – which follows from the first – was the extent to which the psychologists' hypothesis is supported: that the use of unweighted values will give results which are closer than the use of weights or formula. From Table 6, the average difference between scale adjusted MAUI utility is 0.092; the corresponding difference between scale adjusted values is 0.085. This supports the psychologists' hypothesis.

The validity of QALYs as a method for combining the quality and length of life depends upon the adoption of the correct scale: one where a 10 percent change in the quality of life is

deemed equally valuable as a 10 percent increase in the length of life. In the present study the need to employ the correct scale was circumvented. As the purpose was to explain differences between utilities it was sufficient to adopt a single common scale. This leaves unresolved the scale which should be adopted if unweighted scores were to be converted into estimates of utility with the type of linear transformation employed here. The issue is not, however, conceptually complex. A linear transformation may be achieved if two points (V_i, U_i) are known. Since one of these points is necessarily the scale ceiling (1.0, 1.0) the transformation requires, in effect a single point.

A caveat to the present results is that the effect of measurement error – the inconsistent and erroneous completion of two questionnaires – will result in a larger apparent effect of the descriptive systems. The problem is difficult to circumvent as survey respondents are fallible. However it is unlikely to have had a large impact. The MIC data was subject to eight separate edit procedures to delete inconsistent results. These were based upon the comparison of repeated and similar questions and resulted in the removal of 17 percent of respondents from the database before analyses commenced. Remaining inconsistencies are unlikely to explain the magnitude of the effects identified here. A more plausible explanation is that the effect is a correct reflection of the very significant differences in the descriptive systems which are apparent from the casual comparison of the instruments.

A final caveat to the results is that they are necessarily based upon particular published utility formulae. While the effect of the descriptive systems is independent of the utility weighting both the scale and micro utility effects could vary substantially with a change in the utility formula.

Conclusions

The present study suggests that, after allowing for scale, the formula which derive numerical utility scores for the main MAUI have contributed to, not diminished, differences between the utilities predicted by the instruments. This is consistent with the psychologists' hypothesis that the use of weights may reduce, not enhance the validity of measurement.

To our knowledge this is the first study to investigate this issue in the health economics literature. Consequently, any conclusions must be tentative. However the results suggest that

superior instruments might be obtained by a simple adjustment to unweighted scores. This leaves unanswered the question of the appropriate scale to which score should be mapped. The question is fundamental as it determines the trade-off between the quality and length of life implied by the results from MAU instruments. However, the scale may be determined by a single observation which is a less arduous and hazardous task than the creation of a full utility algorithm.

A significant body of research has sought to increase the validity of utility measurement by refining the methods used for eliciting utilities, or by deriving utilities from nationally representative samples. Results in the present paper suggest that such research is unlikely to reconcile the inconsistencies in the utilities predicted by MAUI. Utility weights are shown to be important, accounting for 34 percent of the difference between instrument scores. But their impact is primarily via a scale effect: different utility formula use different scales for the calibration of utility and these account for 30.3 of the 34.0 percent difference between utilities attributable to utility weights. After adjusting for this, the residual effect of different formula – the 'micro utility effect' – is relatively small. This implies that there is little scope for reconciling the numerical values obtained from different instruments by achieving greater precision in the relative values assigned to items. The dominant determinant of the difference between utilities is the difference between descriptive systems. A necessary condition for achieving comparability between utilities, quality adjusted life years and, therefore, the results of cost utility analyses is the use of instruments with comparable descriptive systems or the adjustment of results to take account of structural and scale differences.

References

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

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

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

Feeny, D., Furlong, W., Torrance, G., Goldsmith, C., Zhu, Z., DePauw, S., et al. (2002). Multi attribute and single attribute utility functions for the Health Utilities Index Mark 3 System. *Medical Care*, 40, 113-128.

Gigerenzer, G., & Todd, P.M. (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.

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. In M.D. Dunnett (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally.

Rabin, R., Oemar, M., Oppe, M., Janssen, B., & Herdman, M. (2011). *EQ-5D-5L User Guide: Basic information on how to use the EQ-5D-5L instrument*. Rotterdam: EuroQoL Group, http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/UserGuide_EQ-5D-5L.pdf.

Richardson, J., Iezzi, A., Khan, M.A., & Maxwell, A. (2012). *Cross-national comparison of twelve quality of life instruments: MIC Paper 1: Background, questions, instruments, Research Paper 76*. Melbourne: Centre for Health Economics, Monash University http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper76.pdf [accessed 29 July 2013].

Richardson, J., Khan, M.A., Iezzi, A., & Maxwell, A. (2012b-g). *Cross-national comparison of twelve quality of life instruments, Research Papers 78, 80-83, 85. MIC Report: 2: Australia; 3: UK; 4: USA; 5: Canada; 6: Norway; 7: Germany*: Centre for Health Economics, Monash University, http://www.buseco.monash.edu.au/centres/che/che-publications.html [Accessed 25 January 2013].

Richardson, J., Khan, M.A., Iezzi, A., & Maxwell, A. (2014a). Comparing and explaining differences in the content, sensitivity and magnitude of incremental utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB and AQoL-8D multi attribute utility instruments'. *Medical Decision Making*, doi:10.1177/0272989X14543107

Richardson, J., Sinha, K., Iezzi, A., & Khan, M.A. (2014b). Modelling utility weights for the Assessment of Quality of Life (AQoL) 8D. *Quality of Life Research,* DOI: 10.1007/s11136-014-0686-8.

Sintonen, H., & Pekurinen, M. (1993). A fifteen-dimensional measure of health related quality of life (15D) and its applications. In S. Walker, & R. Rosser (Eds.), *Quality of Life Assessment*. Dordrecht: Kluwer Academic Publishers.

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

van Hout, B., Janssen, M.F., Feng, Y., Kohlmann, T., Busschbach, J., Golicki, D., et al. (2012). Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health,* 15, 708-715.

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

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

Table 1 Comparison of the dimensions and content of five MAU instruments

	Dimension	EQ-5D-5L	SF-6D	HUI 3	15D	AQoL-8D
Physical	Physical Ability/Mobility/ Vitality/Coping/Control	*	*	**	**	**
	Bodily Function/Self Care	*			***	*
	Pain/Discomfort	*	*	*	*	**
	Senses			**	**	**
	Usual Activities/Work	*	*		*	**
	Communication			*	*	*
Psycho-social	Sleeping				*	**
	Depression/Anxiety/Anger	*	*	*	***	*******
	General Satisfaction					****
	Self-esteem					*****
	Cognition/Memory Ability			*		
	Social Function/Relationships		*			*****
	(Family) Role		*			*
	Intimacy/Sexual Relationships				*	*
	Total items/symptoms	5	6	8	15	35
	Health states described	3125	18,000	972000	$3.1x10^{10}$	$2.4x10^{23}$

Table 2 Respondents Characteristics

Country	Composition of Final Sample																	
	Public (%)							Patient (%)							Education			Total (n)
	18-24	25-34	35-44	45-54	55-64	65+	Male	18-24	25-34	35-44	45-54	55-64	65+	Male	High school	Diploma or certificate or trade	University	
Australia	11.3	18.1	18.9	18.5	14.7	18.5	46.4	2.1	8.0	10.3	19.5	32.6	27.5	50.4	35.8	35.1	29.1	1430
USA	10.3	17.8	18.1	20.2	16.2	17.4	45.2	4.8	8.8	13.1	25.0	25.5	22.8	36.4	36.1	29.3	34.6	1460
UK	11.4	15.4	20.1	18.1	14.4	20.5	47.7	7.1	12.7	9.7	16.4	29.0	25.1	51.4	38.1	30.2	31.7	1356
Canada	12.8	18.3	16.2	20.1	16.8	15.9	47.3	5.8	15.1	18.0	19.1	27.3	14.8	34.8	29.2	47.6	23.2	1330
Norway	12.8	16.0	16.7	18.4	15.6	20.5	50.3	6.2	8.2	10.2	16.8	26.0	32.6	63.6	28.0	48.5	23.5	1177
Germany	6.5	20.0	18.5	23.1	17.7	14.2	50.4	5.2	8.3	17.5	31.4	24.4	13.2	54.2	19.6	55.0	25.4	1269
Total	11.0	17.6	18.0	19.7	15.9	17.8	47.8	5.1	10.1	13.1	21.4	27.6	22.6	48.0	31.4	40.4	28.2	8022

Table 3 Summary statistics for the five multi attribute utility instruments (n=8,022)

	Utility					Values			Correlation ρ (U,V)
	Mean	SD	Range	U=1.00 (%)	U<0.4 (%)	Mean	SD	Range	
EQ-5D	0.74	0.23	1.51	19.10	8.90	0.74	0.23	1.30	0.95
SF-6D	0.71	0.14	0.70	1.30	1.30	0.71	0.14	0.62	0.89
HUI 3	0.71	0.27	1.34	7.10	13.90	0.71	0.27	2.10	0.95
15D	0.85	0.13	0.75	6.90	0.30	0.85	0.13	0.67	0.99
AQoL-8D	0.68	0.22	0.90	0.30	14.70	0.68	0.22	1.32	0.98

Table 4 GMS regression of U_i on U_j and U_i on V_j n=8,022

$U_i = a + bU_j$ (equation 3)	R^2	$U_i = a + bV_j$ (equation 4)	R^2
EQ-5D = -0.14 + 1.24 SF-6D	0.57	EQ-5D = -0.20 + 1.32 SF-6D	0.70
EQ-5D = 0.26 + 0.68 HUI 3	0.64	EQ-5D = 0.28 + 0.64 HUI 3	0.62
EQ-5D = -0.50 + 1.45 15D	0.67	EQ-5D = -0.50 + 1.46 15D	0.74
EQ-5D = 0.22 + 0.76 AQoL-8D	0.57	EQ-5D = 0.21 + 0.77 AQoL-8D	0.62
SF-6D = 0.44 + 0.37 HUI 3	0.53	SF-6D = 0.37 + 0.47 HUI 3	0.53
SF-6D = 0.0 + 0.81 15D	0.62	SF-6D = -0.02 + 0.86 15D	0.66
SF-6D = 0.37 + 0.49 AQoL-8D	0.65	SF-6D = 0.38 + 0.49 AQoL-8D	0.61
HUI 3 = -0.77 + 1.75 15D	0.70	HUI 3 = -0.78 + 1.76 15D	0.68
HUI 3 = 0.07 + 0.95 AQoL-8D	0.64	HUI 3 = 0.06 + 0.96 AQoL-8D	0.57
15D = 0.53 + 0.47 AQoL-8D	0.70	15D = 0.53 + 0.48 AQoL-8D	0.75

Table 5 Regression of scale free difference between utilities and difference between values

$[U_i-U_j^*]$ on $[V_i^*)-V_j^*]$ (n = 8,022)

MAU	Pair	Y=a+bX(1)			MAU	Pair	Y=a+bX(1)		
MAU_i	MAU_j	a	b	R^2	MAU_i	MAU_j	a	b	R^2
EQ-5D	SF-6D	0.00	0.83	0.52	SF-6D	15D	0.01	1.05	0.45
EQ-5D	HUI 3	0.00	0.97	0.64	SF-6D	AQoL-8D	0.00	0.94	0.48
EQ-5D	15D	0.00	1.12	0.61	HUI 3	15D	0.00	0.98	0.62
EQ-5D	AQoL-8D	0.00	1.06	0.69	HUI 3	AQoL-8D	0.00	0.92	0.69
SF-6D	HUI 3	0.00	1.00	0.50	15D	AQoL-8D	0.00	1.10	0.85

(1) $Y = [U_i-U_j^*]$; $X = [V_i^*-V_j^*]$

Table 6 Decomposition of (U_i-U_j)

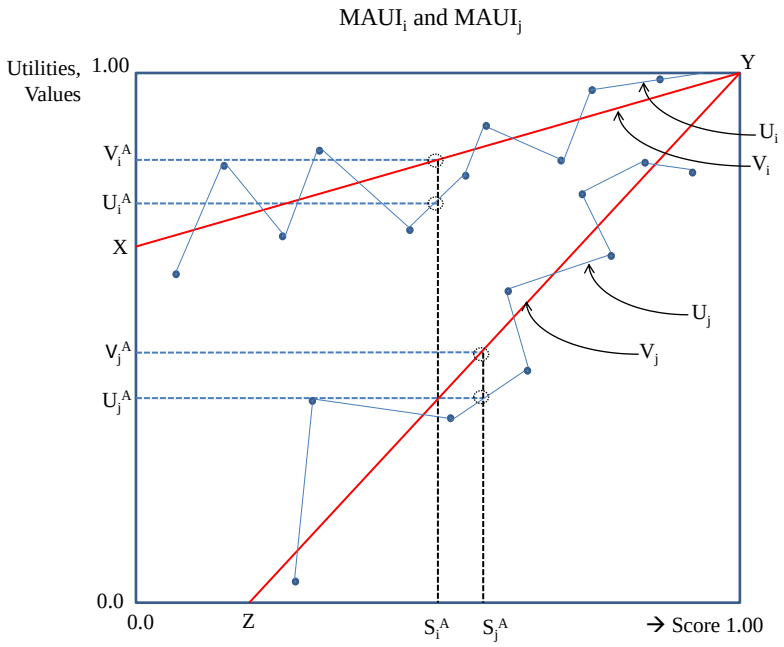
	Absolute Differences					Percent of (U_i-U_j)		
Pairwise comparison[(1)]	**Utility (U_i-U_j)**	**Scale free diff in utility [Ui-U_j^*]**	**Scale effect [A-B]**	**Desc system [V_i-V_j^*]**	**Micro utility [B-D]**	**Scale effect**	**Desc system**	**Micro utility**
	A	B	C	D	E	(C/A)*100	(D/A)*100	(E/A)*100
EQ-5D, SF	0.116	0.112	0.004	0.089	0.023	3.5	76.72	19.8
EQ-5D, HUI	0.117	0.101	0.016	0.101	0.001	13.7	85.5	0.8
EQ-5D, 15D	0.130	0.097	0.033	0.083	0.013	25.7	64.3	10.0
EQ-5D, AQoL	0.130	0.112	0.018	0.105	0.007	13.9	80.8	5.3
SF, HUI	0.146	0.078	0.069	0.075	0.003	47.0	50.9	2.1
SF, 15D	0.144	0.069	0.075	0.062	0.007	52.1	43.0	4.9
SF, AQoL	0.114	0.065	0.049	0.067	-0.002	43.0	58.8	-1.8
HUI, 15D	0.154	0.108	0.046	0.110	-0.002	29.9	71.4	-1.30
HUI, AQoL	0.125	0.120	0.005	0.127	-0.007	4.0	101.6	-5.60
15D, AQoL	0.175	0.053	0.122	0.048	0.005	69.7	27.4	2.9
Average	0.135	0.092	0.043	0.085	0.007[(2)]	30.3	66.0	3.7

(1) SF=SF-6D; HUI = HUI 3; AQoL =AQoL-8D

(2) Average of absolute values

[5]

Figure 1 Hypothetical utilities, U, values, V, and scores, S



Box 1 Definitions

S_i	=	Unweighted score from MAU$_i$
U_i	=	Utility predicted by MAU$_i$ using the published algorithm
U_j (u_i)	=	U_j predicted by MAU$_j$ rotated to the scale of U_i using a linear transformation
V_i	=	Value obtained from the score, S_i, of MAU$_i$ rotated to the scale of U_i
V_j (u_i)	=	Value obtained from the score, S_j, rotated to the scale of U_i