CENTRE FOR HEALTH
PROGRAM EVALUATION

WORKING PAPER 113

Negative Utility Scores and Evaluating the AQoL All Worst Health State

Professor Jeff Richardson
Director, Health Economics Unit, Centre for Health Program Evaluation
Monash University

Dr Graeme Hawthorne
Deputy Director, Program Evaluation Unit, Centre for Health Program Evaluation
The University of Melbourne

June, 2001
ISSN 1325 0663
ISBN 1 875677 82 8

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two indepedent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au
Web Address http://ariel.unimelb.edu.au/chpe

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE is supported by Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR'S ACKNOWLEDGMENTS

The research reported in this working paper has been supported by a project grant from the Victorian Health Promotion Foundation.

Graeme Hawthorne's position at the Centre for Health Program Evaluation at the University of Melbourne is funded by the Victorian Consortium for Public Health.

Table of Contents

Postscript

Following the validation study in 1999 the AQoL utility scoring algorithm was modified and Dimension 1: Illness was removed. Utility scores for the AQoL are now computed by setting the disutility value of the Illness dimension to '0.00' (or the utility value to '1.00'). This results in a modified instrument which still produces scores in the range 0.00–1.00. Results from the validation study indicate that this performs well as a utility instrument. The Illness dimension has been found to be a useful part of a health profile with independent predictive power.

For details see Hawthorne *et al*, **Construction and Utility Scaling of the Assessment of Quality of Life (AQoL) Instrument**. Melbourne: CHPE Working Paper 101 (Hawthorne, Richardson et al. 2000).

Abstract

This paper is concerned with two issues. One is a general problem for the evaluation of very poor health states and the other, while arising from the scaling of the *Assessment of Quality of Life* (AQoL) instrument is a more general problem associated with the use of multiplicative multi-attribute instruments. These problems are:

(i) the treatment of negative utility values;

(ii) the utility score on a 'life-death scale' of the 'all worst' health state described by an instrument (in this particular paper in reference to the AQoL, although the issue is pertinent to all such instruments).

The first problem arises from the fact that there is no lower limit to the negative utility scores implied by responses in a conventional time trade-off (TTO) interview. A respondent who indicates that they would not accept any time in a health state worse than death, even when this was followed by full health, implies a utility score of minus infinity for this health state. While not discussed in this paper, the same outcome is obtained from the standard gamble when a respondent refuses to contemplate a health state at any finite probability. A score of minus infinity or even a very large negative utility score has no meaning and such responses must be transformed into a lower, *albeit* negative, score.

The second problem is that the disutility values generated by a multiplicative model — as used in the AQoL — vary from 0.00 to 1.00. These 'model utilities' — utilities measured in 'model space' — must be rescaled so that they represent utilities on a life-death scale where, following convention, a utility of 1.00 corresponds with 'full' health, and a utility score of 0.00 corresponds with death. In principle it is easy to rescale the model scores. 'Full health' has a common numerical value — 1.00 — on both the model and the life-death scales. The transformation then only requires information on the correspondence between one other point on the two scales. The simplest such point is the instrument 'all worst' (the health state described when each item of the instrument is at its worst level). However scaling this point may require respondents to visualize a particularly complex and unusual state. For example, the AQoL 'all worst' is a health state with 12 dimensions. However, the *raison d'être* of the decomposed, multi-attribute approach to health state measurement is the avoidance of the need to carry out such a cognitively complex task. The likelihood of error is further increased when the instrument's all worst health state is used to establish the nexus, if — as with the AQoL — the all worst health state is close to death and respondents have never experienced a health state so bad that death was equally (un)desirable.

Procedures adopted for the scaling of the AQoL are outlined and discussed. It is concluded that both of the above problems are quantitatively significant and have received too little discussion in the literature.

Readers wishing to understand the scoring of the AQoL are referred to the companion paper, *Utility Weights for the 'Assessment of Quality of Life' (AQoL) Instrument* (Hawthorne, Richardson et al. 2001). Details can also be found in the AQoL user manual, *Using the Assessment of Quality of Life (AQoL) Instrument* (Hawthorne, Richardson et al. 2000).

Negative Utility Scores and Evaluating the AQoL All Worst Health State

1 Introduction

In Cost Utility Analysis (CUA) the cost of a health related intervention is compared with the number of Quality Adjusted Life Years (QALYs) that are obtained because of the intervention. QALYs are usually obtained by multiplying life years by an index of utility where, 1.0 and 0.0 are defined as full[1] health and death respectively. Utility values are obtained either directly by valuing a vignette using one of the standard 'scaling' (calibration) techniques such as the time trade-off (TTO); standard gamble (SG) or rating scale[2] (RS), or indirectly through the use of a multi-attribute utility (MAU) instrument.

MAU instruments typically have two parts. There is a 'descriptive system' (or 'instrument') which is a coherent set of 'items' or statements which describe the different dimensions of health. Each of these has a set of response categories which patients or respondents can use to indicate their own situation with respect to each of the dimensions included in the descriptive system. These responses are then weighted or replaced (referred to as 'scaling' or 'calibration') with values which represent the estimated 'utility' associated with the health state. The weighted or substituted responses are then combined into a single 'utility' score using an algorithm. There are several models available for combining items: *viz.*, an 'additive' model which amounts to a weighted average of the item scores; a 'multiplicative' model, as described below; or an 'econometric' model which employs the regression equation with the best statistical relationship between independently measured health states and the item responses in the corresponding descriptive system. To date only the EQ5D (EuroQoL) has used this last approach (Dolan, Gudex et al. 1995)[3].

To provide a 'valid' utility score — a number which truly represents what it purports to represent (in this case, 'utility') — the numbers produced by each of these techniques must satisfy several demanding conditions. First, they must possess an 'interval' property in the conventional sense. This implies that an increase in the utility score from 0.2 to 0.4 has the same meaning as a move from 0.7 to 0.9. This was described by Richardson as the 'weak interval property'.[4] Second, the utility scores must have a 'strong interval property' (Richardson 1994). This implies that, for example, a 10% increase in the index of utility from, say, 0.7 to 0.77 has exactly the same impact upon utility as a 10% increase in the life years obtained from a project: for example, an increase from 20 to 22 life years. Third, following from both of these requirements, the absolute and not just the relative value of utility numbers must be a valid representation of utility. This third property cannot be directly observed since you cannot demand that people live out their stated preferences.[5]

1 Different researchers describe the top end of the utility scale in different terms. For example, it is referred to as 'Perfect' health by the developers of the HUI3 (Furlong, Feeny et al. 1998) whereas for the EQ5D it is described as 'full' health (Dolan, Gudex et al. 1995). Although the term 'normal' health is commonly used, it is a misnomer. Our data show that the mean utility of a healthy population falls below 1.00. In the case of the AQoL, 'normal' health as defined by a random sample of a healthy population is 0.81. Like the developers of the EQ5D, we describe the AQoL 1.00-value as 'full health'.

2 Also commonly referred to as a 'visual analog scale' (VAS).

3 In principle, it is also possible to fit a non stochastic multi linear model which is even less restrictive than the multiplicative model. Because of the difficulties in the construction of such a model—which requires observations on every possible interaction—it has never been used to model health (Torrance, Furlong et al. 1995).

4 For a discussion of the properties and implications for measurement where interval scaling is not met, see Merbitz, Morris & Grip (Merbitz, Morris et al. 1989).

5 If a person states she is prepared to give up 1/3 of her life to be cured from a health state, you cannot put her in that health state and observe whether or not she does give up 1/3 of her life when she is cured.

An astonishing feature of the CUA literature is that, with the exception of Nord *et al* (Nord, Richardson et al. 1993) and Richardson (Richardson 1994), the latter two properties have received almost no attention despite their almost self evident importance. The possible consequences of their disregard are illustrated in Table 1 which reports the results of four hypothetical programs A, B, C and D. Each of these costs $1,000. The four programs return individuals from health states A, B, C and D to full health. It is assumed that due to measurement error, the true utility (T; column 1) is systematically underestimated[6] as shown in column 2. This implies a measurement scale with an incorrect lower boundary (0.10 below true utility) and which detects only 90% of true increments to utility.

Table 1: Issues in economic and psychometric validation for CUA

Health states & programs (a)	*Utility scores (U)*		*QALYs gained (d)*		*Cost per QALY (e)*		*Patients restored to full health, n, equivalent to saving 1 life (f)*	
	T(b)	*E (c)*	*T*	*E*	*T*	*E*	*T*	*E*
A	0.95	0.755	0.50	2.45	2000	408	20.00	4.10
B	0.80	0.620	2.00	3.80	500	263	5.00	2.60
C	0.40	0.260	6.00	7.40	168	135	2.60	1.40
D	0.20	0.080	8.00	9.20	125	80	1.25	1.05
Column	1	2	3	4	5	6	7	8

Notes:

(a) The health states are hypothetical initial health states for10 years duration before being completely cured at a cost of $1,000. For each health state there is a different treatment program. Thus for Health State A there is Program A, etc.

(b) T = True utility score.

(c) E = Estimated utility score, based on: E = 0.9T – 0.1.

(d) 10(1.00–U).

(e) $1,000/QALY.

(f) $n = {}^{1}/_{(1-U).}$

If this hypothetical utility instrument was subject to a correlation test of validity the measured utilities might be both highly correlated with utilities from other instruments and when compared with true utility scores a correlation coefficient of 1.00 would be obtained. But these findings would be very misleading. Both the QALYs gained and the cost per QALY gained (columns 4 and 6) would deviate significantly from the true QALYs gained and true cost per QALY (columns 3 and 5). From column 8 it can be seen that the instrument would wrongly imply that if Program A could cure more than 4.1 people, then this program should be preferred to another program saving someone's life. However, from column 7, Program A should only be preferred if it cured more than 20 people. Similarly the instrument would wrongly imply that Program B should be preferred to saving a life if it cured more than 2.6 patients. From column 7 it should, in fact, be preferred only if it cured more than 5 patients.

This problem is not simply hypothetical. Nord *et al* demonstrated that the original Quality of Wellbeing (QWB) scale implied the superiority of curing 6 people from pimples or 5 people from a headache to the saving of a single life; and that the Health Utility Index 1 (HUI1) implied the superiority of saving 9 people and 5 people from health states in which they needed a hearing aid or mechanical aids to get around respectively (Nord, Richardson et al. 1993).

[6] This is in relation to either ceiling affects or end-aversion. If these exist for 'full' health states at either the TTO-level when weights are being determined or at the item-completion level then true utility will be underestimated.

The policy implications of poor measurement are equally serious. If a health scheme was prepared to purchase QALYs at a price between $41 and $200 ($408 and $2,000 for 10 QALYs) it would give priority to Program A over saving a life for 10 years. If it paid between $26.30 and $50 it would favor Program B rather than saving a life for the same cost.

The purpose of the present paper is to investigate two related problems which may lead to the type of error illustrated above, and to use the results of the analysis to determine the scoring system for the *Assessment of Quality of Life* (AQoL) multi-attribute utility instrument.[7] The first of these problems is the treatment of negative scores produced by the use of the TTO procedure, and the need for their transformation before combination with positive utility scores.

The second problem is the determination of the (dis)utility value of the AQoL all worst health state, W: the health state described by the worst item response from each item (see Appendix 1)[8]. This is related to the first problem because of the large number of interview respondents (76%) who valued W as being worse than death. Consequently, error in the measurement of negative utilities will lead to error in the measurement of W. This is compounded by the complexity of the cognitive task. It is unlikely that many of our respondents could easily digest the fifteen pieces of information contained in W (one for each item) and determine a utility score which satisfies the stringent criteria for validity discussed above.

This implies a more general problem. There are compelling reasons for using a multiplicative model (Richardson and Hawthorne 1998), but this requires the estimation of an 'exchange rate' between the model scale where the endpoints are defined as 0.0 (instrument all-worst) and 1.0 (instrument all-best) and a scale on which 1.0 and 0.0 represent full health and death respectively. This exchange rate may be estimated one of two ways. A health state other than the instrument all-worst may be evaluated during the scaling survey and a 'bridge' established between scores predicted by the multiplicative model and the score obtained on the life/death scale. This has the advantage that a relatively simple health state may be selected which respondents can easily visualize and may even have experienced. However as will be seen, this approach is problematic since the predicted score depends, itself, upon the final exchange rate; that is, the health state value is 'endogenous'. Alternatively, the instrument all-worst health state may be evaluated. This avoids the problem of endogenuity. However, as noted above, when the instrument has as many items as the AQoL it is unlikely that respondents to the scaling survey will have the cognitive capacity to fully appreciate the implications of the health state. This casts doubt upon the validity of the exchange rate. Unless an acceptable solution can be found to this problem, then either unsatisfactory models must be used to combine instrument dimensions or an instrument's descriptive system must be excessively simple.

The present paper describes these problems more fully and outlines the procedures used to overcome them during the scaling of the AQoL instrument.

---

7 A description of the AQoL and its construction is provided elsewhere (Hawthorne, Richardson et al. 1997; Hawthorne, Richardson et al. 1999; Hawthorne, Richardson et al. 2000).

8 W = A person who uses five or more medicinal drugs regularly, constantly takes medicines or uses a medical aid, is dependent upon regular medical treatment, needs daily help with most or all personal care tasks, needs daily help with most or all household tasks, cannot get around either the community or his/her home by his/herself, who has no close and warm relationships, is socially isolated and feels lonely, who cannot carry out any part of his/her family role, who only sees general shapes or is blind, who hears very little, who cannot adequately communicate with others, who sleeps in short bursts and is awake most of the night, who is extremely anxious, worried or depressed, and who suffers unbearable pain. In the notation adopted below this can be described as (444, 444, 444, 444, 444).

The survey results reported in this paper come from the survey carried out to elicit the utility weights for the AQoL instrument. Although the survey is not described here, details of it can be found in the companion paper, *Utility Weights for the 'Assessment of Quality of Life' (AQoL) Instrument* (Hawthorne, Richardson et al. 2001).

2 Negative Utilities

Negative utilities were first reported by Rosser & Kind (1978) and subsequently by other researchers (Rosser and Kind 1978). In an early discussion, Torrance simply noted that the measurement of states worse than death was 'still at a very early, primitive stage' (Torrance, Boyle et al. 1982, p1083). In his seminal 1986 review of utility measurement theory and practice, Torrance described the technique for measuring negative utility and reported that it results in numerical values as low as minus infinity. He suggested these scores should be constrained to –1.00 on the grounds that this achieves symmetry: if the greatest positive utility is +1.00 then the lowest negative score should be –1.00 (Torrance 1986). This advice was accepted in the scaling of the EuroQoL (Williams 1995b).

Like the EuroQoL, the AQoL was scaled using the TTO and the standard method was adopted in the first stage of the estimation of negative values. When respondents indicated that they would rather be dead than be in a particular health state for 10 years prior to death they were offered a second choice (which was illustrated using a TTO 'slide board'). With this they were asked whether they preferred the option of immediate death or a period of time, x, in the health state, h, followed by the remainder of the 10-year period, 10-x, in full health as defined by the AQoL 'all best' health state (all item responses set at their highest level). The number of years, x, was varied until the respondent indicated that the two options were equally attractive (or unattractive!). The value of x was recorded and subsequently the implied utility of the health state, V, was calculated from the equation:[9]

$$V = \frac{-(10-x)}{x}$$

Equation 1

From Equation 1, as x varies from 10.00 to 0.00, the value of V varies from 0.00 to $-\infty$.

Table 2 shows the distribution of negative scores (raw data) for the five AQoL dimension all worst scores (where the three items in each dimension assume their worst level and where other AQoL values are at their best). It also reports the AQoL 'all worst' value, W. The distribution for this final health state indicated that 76% of respondents believed that 10 years in health state W would be worse than immediate death.

[9] At this point, x years of the health state, V, plus (10-x) years of full years, are equal to death. With 'full health' and 'death' set equal to 1.00 and 0.00 respectively, $x \cdot V + (10-x) \cdot 1 = 0$, therefore $V = -(10-x)/x$.

Table 2: Distribution of negative scores (worse than death evaluations (a))

		Dimensions (b)					*AQoL*
		1	***2***	***3***	***4***	***5***	
Numbers (%)	Positive scores	125 (74%)	99 (59%)	98 (58%)	75 (45%)	62 (38%)	36 (24%)
	Negative scores	43 (26%)	69 (41%)	70 (42%)	92 (55%)	103 (62%)	113 (76%)
Distribution of negative scores for worst health states (c)(d)							
	–10.00–9.01	1	4	5	4	5	21
	–9.00–8.01	3	6	8	11	25	36
	–8.00–7.01	1	1	6	12	9	15
	–7.00–6.01	3	7	2	4	10	8
	–6.00–5.01	0	4	3	5	8	6
	–5.00–4.01	7	21	13	22	21	13
	–4.00–3.01	5	5	4	4	3	3
	–3.00–2.01	9	4	9	4	5	3
	–2.00–1.01	5	6	5	6	7	0
	–1.00-0.01	3	4	5	9	5	4
	0.00	6	7	10	11	5	4

Notes:

a = When a respondent indicated that the health state was evaluated as 'worse than death', he/she was asked how many years (x) in the health state followed by full health for the remainder of a 10-year period (10–x) and then death would be equivalent to immediate death for the entire 10-year period (followed by death!).

b = 1: Illness; 2: Independent Living; 3: Social Relationships; 4: Physical Senses; 5: Psychological Wellbeing

c = Dimension worst health states were where all items within the dimension were set at their worst health state, ie. 444. For the AQoL the 'all-worst health state' ('W') was defined as (444, 444, 444, 444, 444).

d = Values are raw data, grouped and constrained between 0.00 and –10.00. When transposed onto a Full-health (+1.00) — Death (0.00) scale, the following conversions apply: –10.00 = 0.00, –9.00 = –1.00 etc. Thus a person who indicated –8.00 in the table was indicating that he/she would be willing to spend 2 years in the health state in exchange for 8 years of full health following the 2 years of illness.

As noted by Torrance and the EQ5D (EuroQoL) team, large negative scores have little meaning. They may provide an ordinal index of the ranking of health states but it is implausible to assign an interval property to the numerical values and, for this reason, it is necessary to transform the negative scores and to constrain them in some way. This raises two separate problems which have not been discussed in the literature. Both concern the nature of the transformation between the disvalue scores, V (calculated from Equation 1), and the disutility scores, DU, which we accept as an appropriate representation of the strength of a person's dislike of the health state and which may be used in the calculation of average utility (U) scores. These problems are: (a) that it is necessary to determine the upper limit to DU; and that (b) the transformation path between the value of death and this limiting disvalue score must be determined.

Maximum disutility

Disvalue may be transformed so that the negative 'value' scores between zero and minus infinity are constrained to disutility values of 1.00 and $c = 1 + m$ respectively, where m is the maximum possible disutility. A function that will achieve this transformation is [10]:

$$DU = c + \left(\frac{1}{\left(V - \frac{1}{m} \right)} \right)$$

Equation 2

When the score for the untransformed value of a health state (V), equals zero (death) then $DU = 1.00$. When V is equal to minus infinity, $DU = 1 + m$, where m is the magnitude of the maximum negative utility (ie. $U = -m$).

The choice of m has a significant impact upon 'observed' disutilities, ie. the disutility value that is obtained when V is negative and DU is calculated from Equation 2. This is illustrated in Table 3 where the 'observed' disutilities for five multi-attribute health states are reported when they are calculated with different values for the maximum permissible disutility.

Table 3: Observed disutilities of selected multi-attribute health states as maximum DU varies using illustrative transformations (a)

Health State	*Health state definitions*	*Maximum DU*					
		2.0	***1.5***	***1.4***	***1.3***	***1.2***	***1.1***
A	(444, 444, 111, 111, 111)	1.14	0.959	0.929	0.901	0.876	0.855
B	(111, 111, 444, 444, 111)	1.33	1.092	1.047	1.005	0.967	0.934
C	(333, 111, 333, 111, 432)	0.84	0.750	0.735	0.721	0.709	0.699
D	(111, 333, 111, 313, 432)	0.77	0.693	0.680	0.668	0.657	0.649
W	(444, 444, 444, 444, 444)	1.53	1.205	1.143	1.084	1.028	0.979

Notes:

a = The function for *DU* used here for illustrative purposes is equivalent to the function used later when $m = 1$ and $n = 1$. As this is not the value of n finally adopted, data in Table 3 do not correspond with later data.

n = 146

For the disutilities reported in Table 3, m assumes values between 1.00 and 0.10; ie. $1+m$ corresponds with 2.00 and 1.10 respectively. For all five states the choice of this lower boundary has a significant quantitative affect upon the 'observed' disutility. In three cases the value of the state switches from better than to worse than death (see health states A, B and C). For the AQoL all worst health state, W, increasing m from 0.10 to 1.00 (as recommended by Torrance and used by the EQ5D (Williams 1995a; Williams 1995b)) increases the disutility of W by 56%[11]. If, as argued below, a disutility boundary of 2.00 is too large, then adopting this lower boundary will significantly lower the utility scores predicted by an MAU instrument and will bias evaluation studies in favor of those curing serious health states.[12]

[10] See the footnote to Table 3.

[11] The change in values is from 0.979 to 1.530; see health state *W* in Table 3.

[12] For example, consider two interventions A and B which return patients to full health from health states with true utility values of 0.7 and 0.9 respectively and which, thereby, increase utility by 0.3 and 0.1 respectively. If the bias in an instrument doubled disutilities then the apparent increase in utility from the two interventions would be 0.6 and 0.2 respectively. The difference in the apparent effectiveness of the two interventions would increase from 0.2 (0.9-0.7) to 0.4 (0.6-0.2).

Despite the recommendation by Torrance, there are neither theoretical nor empirical reasons for accepting a lower boundary of DU = 2.00. The appeal to 'symmetry' — the argument that the upper and lower boundaries should have the same absolute value — cannot be supported except by the generally favorable connotations of the word 'symmetry'. But the symmetry is only superficial and results in an asymmetry in the logic of the disutility calculation. With a ten year life expectancy, positive utility scores are calculated from the individual's statement that $10 \cdot U = x$ years, where U is the utility of the health state. From this, $U = (x/10)$ and U will rise by 0.10 each time x rises by 1.00. That is, each time U rises by 0.10 the increase in ten year utility is equal to the utility gained by one full year of good health.

In the negative range of 'utilities' there is no similar logic. Each time x increases by one year the implied disutility (V; Equation 1) changes by more than the utility associated with twelve months in good health. This is because a one year increase in good health is associated with one less year of poor health. For example, when x = 10, 9, 8, 7 etc. the value of V is 0/10; 1/9; 2/8; 3/7 etc. or 0; –0.11; –0.25; –0.46 etc. These untransformed values are not consistent with the logic in the positive range of the scale. Thus, scores of –1/9, –2/8, –3/7 etc. mean that an individual believes that 1, 2, 3 etc. years in the health state is so bad that they would be prepared to sacrifice 9, 8, 7 etc. years to avoid being in the health state for these periods of time. The 'utility' implied by this information is then arbitrarily transformed to a figure between 0.00 and –1.00 whose literal meaning has no simple relationship with the utility of a twelve month period of good health. It is in this sense that there is no symmetry in the meaning of positive and negative disutility scores and the selection of the maximum disutility is necessarily arbitrary.

At best, the symmetry argument reflects the judgement that each person *should* be allowed to affect total utility by the same amount; if one person can add 1.0 units of utility then a second person should be permitted to subtract the same quantum. But this is an appeal to individual rights and the *prima facie* argument that all people *should* be equal in this particular respect. If accepted, this argument results in a total utility score which combines positive utilities, which reflect a psychological quantity, and negative scores which reflect a rather vague notion of a quantified democratic right. This 'right' is highly contentious. It is the assertion that one person may fully obliterate the benefits of life to a second person and it is far from obvious that this powerful value judgement would receive broad endorsement. More to the point here, the combination of numerical scores which arise from two different sets of consideration (*viz.,* intensity of feeling and index of democratic rights) results in an overall 'utility' which has little if any meaning.

As the *sine qua non* of the QALY is its ability to provide an exchange rate between the quantity and quality of life, this inability to provide a sensible, literal meaning for the numerical values in the negative utility range represents a serious threat to the validity of the numerical scores. The literal interpretation given above clearly results in invalid scores. Values of x of 1, 1/2, 1/10, 1/50 correspond with utility scores of –9.0; –99.0; –49,900 and it is clearly misleading to place any psychological interpretation on a health state so bad that 49,900 years of full health would be sacrificed to avoid it for 1 year. Untransformed negatives scores therefore cannot have either the strong or weak interval property. The necessary transformation which converts these ordinal indices into scores which, hopefully, have a cardinal property must, at present, be based upon a subjective judgement.

With the AQoL, the upper value (full health: 1.00) was based upon two considerations, the first theoretical and the second empirical. The theoretical argument arose from a reconsideration of the meaning of 'utility'. While this term has been used by economists in several different ways (Richardson 1994) the general and most useful interpretation is that 'utility' represents the *intensity* of a person's preferences[13]. With this interpretation the literal meaning of a utility score of –1.00 is that the change in the intensity of a person's preference between a utility of 0.00 and –1.00 is just as great as the change in the intensity between the utility of 1.00 (full health) and the utility of 0.00 (death). For a health state to be equivalent to death — to be sufficiently bad that a person will override their most basic instinct to live — it must be truly awful and certainly close to the limits of a person's endurance.[14] The interval between full health and this terrible health state is correspondingly large. It is simply not plausible that a human being is capable of experiencing an intensity of preference in any meaningful psychological sense that is so great that it could further reduce intensity by the same quantitatively enormous interval. Restating this, it is implausible that there could be a health state so terrible that in any psychological sense it could further reduce the quality of life by a quantum equivalent to the interval between full health and death. This implies that disutilities in the vicinity of –1.00 have, at best, ordinal meaning.[15]

This argument implicitly uses the criterion which we believe should be explicitly employed to determine maximum disutility. This is that the lower boundary should be determined by the maximum capacity of a person to experience an intensity of feeling and that the numerical value should be determined by the need to preserve the interval property of the utility scale. In the positive utility range it is argued that this property is achieved through a person's capacity to appreciate the value of a life year and their capacity to appreciate the implication of a change in the number of life years. For the reasons just given, this property cannot be preserved in the negative range using the TTO.[16]

To obtain evidence on the order of magnitude of the lower boundary we conducted a survey which, *inter alia*, explored the upper and lower end points of the AQoL utility scale. One hundred and sixteen respondents, randomly chosen from the Victorian community, were asked to locate death on a 100-point visual analog scale which was anchored at 'best imaginable' and 'worst imaginable' health states. The median value of death on this scale was 10.00 (with a mean score of 13.6). Recalibrating this scale so that death assumed a value of 0.00 implied a median value for the worst imaginable health state of –0.11 or a disutility of 1.11 (the mean was –15.7 or a disutility of 1.157).

[13] 'Preference utilitarianism' in philosophy is often interpreted somewhat differently. In this, a person may or may not be aware that the state of the world for which he has a preference has been achieved. Thus, for example, if a person has a preference for the fidelity of his spouse and – with or without the person's knowledge – the spouse is, indeed, faithful then utility will be increased.

[14] For example, consider a person who commits suicide. She has made the evaluation that her current health state is worse than continuing life: she has made the decision she cannot endure life any longer.

[15] It is possible to give a non-psychological (feeling based) meaning to large negative scores; *viz.*, that when faced with the options which result in this score the individual will choose death. Of course, this is a mere restatement of the question used to generate the score. At best, it may accurately reflect a person's revealed behavior if such hypothetical choices became real. However, 'utility' scores then have only ordinal significance: they indicate the rank order of the dislike of the various health states but cannot claim to have an interval property which would justify the addition of negative and positive utilities as suggested by utilitarian principles.

[16] Since this difficulty is intrinsic to the very properties of negative utilities, we do not believe that it can be overcome using another scaling technique.

This value appears to be plausible. It is not inconceivable that individuals can imagine a health state which is 11 percentage points worse than death on a utility scale with interval properties. It is known, however, that rating scales suffer from end-aversion effects. This implies that the numerical value of the worst imaginable health state may have been constrained when compared with the numerical value embodying an interval property based upon a TTO. For this reason we increased the maximum disutility score, arbitrarily and marginally to a numerical value of 1.25 on a 0.00–1.00 disutility scale where 0.00 and 1.00 represent full health and death respectively. This is a significantly lower disutility (higher utility) than the values incorporated in previous studies.

The transformation function

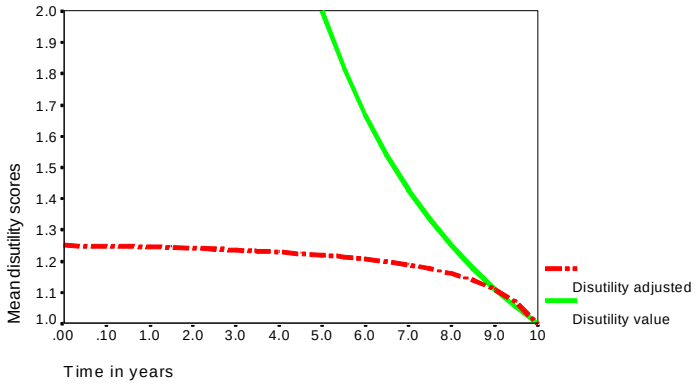
There are an infinite number of transformation functions which constrain disutility between zero and any nominated maximum disutility. One such function is a modification of Equation 2, where c is the lowest permitted value and n determines the transformation function for the negative scores:

$$DUA = c + \frac{1}{nV - \frac{1}{m}}$$

Equation 3

For the reasons discussed below the parameters in Equation 3 were set at c = 1.25 and n = 28.6. Figure 1 illustrates the transition paths for untransformed 'disutility', DU (Equation 2), and the transformation function which uses these parameters.

Figure 1: Unadjusted vs. adjusted disutilities



Notes to Figure 1:

x Years of good health; therefore 10-x = years of poor health.

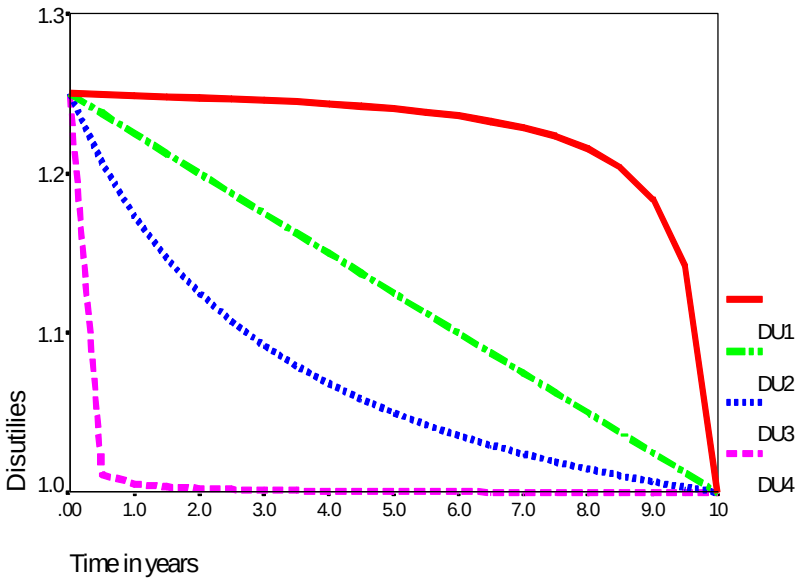
DU = disutility, from $V = -(10 - x)/x$

$$DU = 1.25 + \frac{1}{(nV - 4)}.$$

Where n = 28.6 for reasons discussed later in the text.

Figure 2 illustrates four other transformation functions which correspond with n = 100.00, 4.00, 1.00 and 0.10 respectively. When AQoL scores were computed using the transformation, $DU2$, the health states in Table 3 were a maximum of 4 percentage points greater than when we employed $DU3$.

Figure 2: Four transformation patterns for disutilities



Notes to Figure 2:

Calculated from: $$DUA = c + \frac{1}{nV - \frac{1}{m}}$$

DU1: n = 100
DU2: n = 4
DU3: n = 1
DU4: n = 0.01

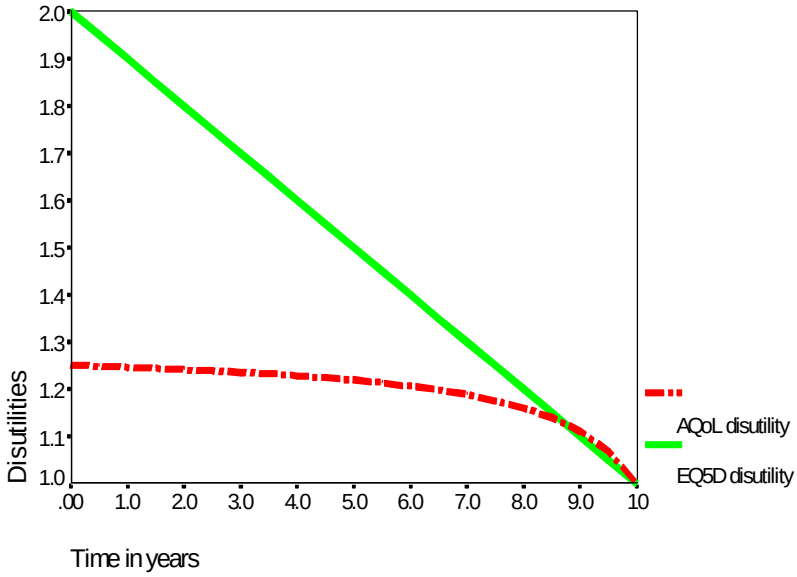
As Figure 2 shows, however, there is no firm guideline to indicate the choice of the transformation function. Our adopted procedure embodies the assumption that untransformed utility is a valid measure of preferences and retains an interval property for half of the time period needed for untransformed utility to reach the maximum disutility of 1.25 and that, thereafter, it follows a smooth transformation path to the maximum disutility. More specifically, we first observed that the unconstrained disutility, DU, reached the maximum disutility (1.25) when x, the number of years obtained during the interview, equaled 8.0 (Figure 1). The transformation function was selected so that it closely tracked unconstrained disutility, DU, until it had reached the mid-point of the interval between 10 and 8 years. At this point — 9 years — we forced the transformation function to intersect the unconstrained disutility function, DU. The transformation function which achieved this was where n = 28.6, giving:

$$DUA = c + \frac{1}{28.6 \cdot V - \frac{1}{m}}$$

Equation 4

This is plotted as the adjusted disutility in Figure 1, and is contrasted with the transition path used by the EQ5D in Figure 3. Subsequent analyses adopted this transformation.

Figure 3: AQoL vs. EQ5D disutility paths



Notes to Figure 3:

Calculated from: $$DUA = c + \frac{1}{nV - \frac{1}{m}}$$

EQ5D disutility path, where c = 2, n = 1 and m = 1.

AQoL disutility path, where c = 1.25, n = 28.6 and m = 0.25.

Table 4 reports the actual observed disutilities from the AQoL weighting survey described above, and shows two sets of multi-attribute health states; *viz.*, each of the AQoL dimension all-worst health states (where the other dimensions assume their best values) and seven other multi-attribute states including the AQoL all-worst, W. The table reports the observed values of these, first, when negative states are determined by the selected transformation discussed above (column B). Second, and for comparison, they are reported in column A for an alternative transformation defined by:

$$DUA = 1.25 + \frac{1}{4V - 4}$$

Equation 5

Where V is the unconstrained value of disutility.

Table 4: Observed disutility of 12 multi-attribute health states from the AQoL using two different transformations of negative utilities

Health States		Observed mean (sd) disutility values	
	AQoL definition	A (1)	B (2)
I Dimension worst state			
Dimension 1	(444, 111, 111, 111, 111)	0.62 ± 0.38	0.64 ± 0.41
Dimension 2	(111, 444, 111, 111, 111)	0.85 ± 0.31	0.86 ± 0.34
Dimension 3	(111, 111, 444, 111, 111)	0.86 ± 0.31	0.89 ± 0.33
Dimension 4	(111, 111, 111, 444, 111)	0.95 ± 0.28	0.97 ± 0.30
Dimension 5	(111, 111, 111, 111, 444)	1.00 ± 0.28	1.04 ± 0.30
II Combination health states			
A	(444, 444, 111, 111, 111)	0.92 ± 0.29	0.95 ± 0.32
B	(111, 111, 444, 444, 111)	1.02 ± 0.24	1.06 ± 0.26
C	(333, 111, 333, 111, 432)	0.73 ± 0.35	0.75 ± 0.37
D	(111, 333, 111, 313, 442)	0.68 ± 0.34	0.69 ± 0.36
E	(141, 411, 114, 114, 114)	1.00 ± 0.23	1.03 ± 0.25
F	(114, 141, 141, 141, 141)	0.90 ± 0.31	0.92 ± 0.33
W	(444, 444, 444, 444, 444)	1.09 ± 0.22	1.12 ± 0.23

Notes:
Data from AQoL weighting survey, n = 162
The data do not correspond with Table 3; see Note (a) to Table 3.

1 = For negative utilities, V, $DU = 1.25 + \left(\frac{1}{4V - 4}\right)$

2 = For negative utilities, V, $DU = 1.25 + \left(\frac{1}{28.6V - 4}\right)$

3 Transforming Predicted Utility to the Life Death Scale

Section 2 above considered the first of the two problems which are the subject of this paper; *viz.*, the adjustment of 'observed' multi-attribute health states when the observations include negative scores from the TTO evaluation.

The second problem concerns the prediction of utility scores on a 'life-death scale' – a scale where full health and death have values of 1.00 and 0.00 respectively – from the 'model scores' produced by the multiplicative model which are constrained to the range (0.00 to -1.00) or a disutility score in the range (1.00 to 0.00).

The two problems overlap. Because actual model scores are substituted with weighted scores which are then combined through use of a multiplicative algorithm, the treatment of negative values influences this transformation.

The general form of the multiplicative disutility function used by the AQoL is given in Equations 6 to 8 (von Winterfeldt and Edwards 1986):

$$DU = \frac{1}{k}\left[\prod_{i=1}^{n}\left[1 + kw_i DU_i\left(x_{ij}\right)\right] - 1\right]$$

Equation 6

$$k = \prod_{i=1}^{n}(1 + kw_i) - 1$$

Equation 7

$$U^* = 1 - DU^*$$

Equation 8

Where:

w_i = weight for dimension i; and

$DU(x_{ij})$ = dimension disutility for item responses, $_j$ (0.00 –1.00 scale).

As noted in our companion paper (Richardson and Hawthorne 1998) this model is significantly more flexible than the simple additive model employed in some MAU instruments (eg. the 15D). When a disutility score of $DU\ (x_{ij}) = 0$ for all dimensions, i, Equation 6 reduces to:

$$DU = \frac{1}{k}(1-1) = 0$$

Equation 9

When $DU\ (x_{ij}) = 1$ the equation reduces to:

$$DU = \frac{1}{k}\left[\prod_{i=1}^{n}(1+kw_i) - 1\right]$$

Equation 10

This is turn, from Equation 7, reduces to:

$$\frac{1}{k}(k) = 1.00$$

Equation 11

Despite the apparent complexity of Equations 6 and 7 the multiplicative model imposes a very simple and specific structure upon preferences between the two extreme values obtained in Equations 9 and 11. This may be seen by setting all disutility scores, $DU_I(x_{ij}) = 1.00$. If global utility, DU, is calibrated so that it assumes a score of 1.00 when dimension scores each have a score of 1.00 then the left hand side of Equation 6 equals 1.00 and Equation 6 becomes Equation 7. Thus Equation 7 simply states that a value of k must be selected such that DU is calibrated to equal 1.00 for the model all-worst score.

The structure of Equation 6 can be readily understood by setting k = – 1.00. This occurs if any dimension has a disutility on the life-death scale equal to the instrument all-worst utility score. With this value Equation 6 reduces to:

$$DU = -\prod_{i=1}^{s}\left(1 - w_i DUx_{ij}\right) + 1$$

As $DU(x_{ij})$ is dimension disutility on a 0.00–1.00 scale where w_i is the maximum disutility of the dimension on the life death scale (LD), $w_i(DU_{ij})$ is dimension disutility on the life death scale.

Thus:

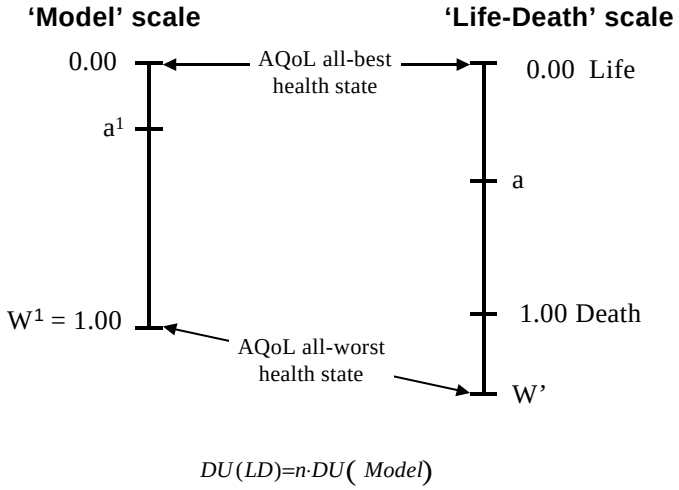
$$DU = 1 - \prod_{i=1}^{5} \left(1 - DU_i^{'}\right) = 1 - \prod_{i=1}^{5} \left(U_i^{'}\right)$$

Where $U_i^{'}$ and $DU_i^{'}$ are dimension *i* utility and disutility on the LD scale. As $U = 1.00 - DU$, then:

$$U = 1 - \{1 - \prod_i (U_i)\}$$
$$U = U_1 U_2 U_3 U_4 U_5$$

Because the multiplicative model is constructed so that predicted utilities will lie in the range (0.00, 1.00), it is necessary to map 'model disutility' into a second scale in which 0.00 and 1.00 represent full health and death respectively. The transformation is illustrated in Figure 4.

Figure 4: Mapping 'model' onto 'life-death' utilities



In principle this task is straightforward when the model involves disutilities (and it is partly for this reason that modeling is usually conducted in terms of disutility scores). Establishing a single 'bridge' or equivalence between any two points and a^1 and a on the 'model' and 'life-death' scales respectively will permit the recalibration of the model utility values using the equation:

$$U(LD) = \left(\frac{a}{a^{'}}\right) U(Model)$$

Equation 12

The AQoL all-worst health state, W, is an obvious choice for calculating the bridge. It has the advantage that W, as an extremely poor health state, will be in the vicinity of death for most people. Asking survey respondents to establish this particular bridge forces them to explicitly consider a life-death decision. If used in Equation 12, $a^1 = 1.0$, $a = W$ and therefore $a/a^1 = W$.

The validity of this approach (or more generally the validity of the values obtained from any model) depends upon the validity of the internal structure of the AQoL and the extent to which the multiplicative model truly represents preferences. Any error in this structure will result in an error in the extrapolated value of W. There are, in addition, two disadvantages with the use of W. First, it involves a consideration of 15 item responses. Even with some simplification of this scenario there is a significant likelihood of cognitive overload. Second, a health state as unpleasant as W may evoke a 'shock horror' effect which could unduly increase the disutility score. A variant of this approach, therefore, would be to ask respondents to select a combination of items which would result in a health state which in their opinion was equivalent to death. There is, however, no reason to believe that this would overcome the 'shock horror' problem as few respondents would have seriously contemplated such health states.

Endogenous estimates

An alternative to establishing the bridge at W or in the vicinity of death is to select an intermediate health state as the bridge. In Figure 4 a^1 and a are the two scores of such an intermediate health state on the model and on the life-death scales respectively. In principle, any such intermediate health state may be used for this purpose and the AQoL all-worst health state predicted from the relationship $W/W^1 = W/1.0 = a/a^1$. From this, $W = a/a^1$.

This intermediate solution, however, has a serious problem which may be seen by setting all dimension utility scores equal to zero, except for dimension 1. This simplifies Equation 6 to:

$$DU_1\,(Model) = \frac{1}{k}\left\{1 + kw_1 DU_1\left(x_j\right) - 1\right\}$$

Equation 13

Which gives:

$$DU\,(Model) = w_1 DU_1\left(x_j\right)$$

Equation 14

Further simplifying by setting $DU_1\left(x_I\right) = 1.00$ (ie. at its worst value) then $DU\left(Model\right) = w_i$.
This states that dimension disutility measured in 'model space' equals the importance weights, w_i, which, by construction, is also equal to the dimension disutility measured in 'life death' space. From Equation 12, when the model scores are transformed into disutility measured on a life death scale:

$$DU\left(LD\right) = \left(\frac{a}{a'}\right) \cdot DU\left(Model\right) = W \cdot DU\left(Model\right)$$

Equation 15

Substituting from the above (ie. $DU\left(Model\right) = w_i$):

$$DU\left(LD\right) = W \cdot w_1$$

Equation 16

This final expression gives an incorrect valuation of disutility on the life death scale, because w_1 is measured directly on the life death scale. Therefore $DU_1\left(LD\right) \neq w_1$. For this reason it is

necessary to replace the weights w_i with adjusted weights w_i/W. When these are substituted in the model the disutility value for the worst health state for dimension 1 becomes:

$$DU\left(LD\right) = W \cdot \frac{w_1}{W} = w_1$$

Equation 17

Which is the (correct) observed value.

In essence, the use of unadjusted weights introduces error because the unadjusted weights are set equal to the dimension disutility values measured in 'life death utility space'. When they are used in the model defined by Equation 6, the resulting disutility index numbers are constrained to the 0.00 –1.00 range 'model space'. However, after transformation from model space to life-death utility space all disutility model numbers are increased by the factor W_i. The dimension all-worst scores (which equal dimension weights) are therefore inflated, erroneously, by W, and must therefore be decreased by a factor of $1/W$.

This revision of the model introduces a complication in the estimation of the AQoL all-worst. While it is still true that $W = a/a^1$, the numerator of this term, *viz.*, the predicted utility for health state a, measured on a life-death scale, is now determined by the final AQoL model which includes the transformation to life-death space and this, in turn, presupposes the value of W. This is a problem of simultaneity which may be solved algebraically: since $W = a/a^1$, $a = Wa^1$ where a^1 is the predicted model score given by Equation 6. Using Equation 13 and replacing w_i by w_i/W, this last expression gives:

$$a = W \cdot \frac{1}{k}\left\{\prod_i\left(1 + k\,\frac{w_i}{W} \cdot DU_1\left(x_{ij}\right)\right) - 1\right\}$$

Equation 18

For the case of three non-zero health states, and simplifying notation, this reduces to:

$$a = -W\,\frac{1}{\mathrm{k}}\left\{\left(1-\frac{d_1k}{W}\right)\left(1-\frac{d_2k}{W}\right)\left(1-\frac{d_3k}{W}\right)-1\right\}$$

Equation 19

Where $d_i = wi \cdot DU\left(x_{ij}\right)$ this reduces to:

$$0 = \left(d_1 + d_2 + d_3 - a'\right)W^2 - \left(d_1d_2 + d_2d_3 + d_1d_3\right)\cdot k \cdot W + \left(d_1d_2d_3\right)\cdot k^2$$

Equation 20

In the case of two non-zero dimensions, $d_3 = 0$, from which

$$W = \frac{d_1d_2 \cdot k}{(d_1 + d_2) - a}$$

Equation 21

The values of W and k for the multi-attribute health states in Table 5 may be obtained by combining each of Equations 19, 20 and 21 respectively with Equation 7. In each case this results in two equations which may be solved simultaneously to obtain the values of W and k. The resulting health state utilities are reported in Table 5.

Table 5: Endogenous estimates of *k* and *W* for selected AQoL health states

State (a)	*Observed definition*	*'Observed'*	*k*	*W*
A	(444, 444, 111, 111, 111)	0.95	-1.00	0.99
B	(111, 111, 444, 444, 111)	1.06	-1.00	1.06
C	(333, 111, 333, 111, 432)	0.75	-1.00	0.91
D	(111, 333, 111, 313, 442)	0.69	-0.99	0.80
E	(141, 411, 114, 114, 114)	1.03	-1.00	1.04
F	(114, 141, 141, 141, 141)	0.92	-0.99	0.98
Mean				0.96

Exogenous estimates

A shortcoming with the simultaneous solution is its sensitivity to a small error in the observations. This is easily seen from Equation 21 when the coefficient a is increased. For example, using AQoL values a 10% increase in a will increase the estimate of W by 23%. For this reason a second set of estimates were derived assuming W to be exogenous and equal to a/a^1 (observed divided by predicted from equation values), with a computed assuming that the dimension weights are equal to the observed dimension all worst health states divided by the observed value of W, W = 1.12. The consequence of this assumption is that the implied values of W will be somewhat higher than true values. These estimates are shown in Table 6.

Table 6: Exogenous estimates of selected AQoL health states

State (a)	*Observed Utility (b)*	*Predicted (Model) Utility (a^1)(^{2})*	*Implied W (a/a^1)*
A	0.95	0.91	1.06
B	1.06	0.97	1.09
C	0.75	0.74	1.02
D	0.69	0.72	0.97
E	1.03	0.98	1.05
F	0.92	0.90	1.02
W	1.12	1.00	1.12
Mean			1.04

Notes:
a = See Table 5 for AQoL definitions
b = From utility survey
W is set equal to 1.124, as described in the text

4 Selection of the All Worst Utility, W, and the Final Model

From Tables 5 and 6 there are a number of estimates of W ranging from 0.81 to 1.12. The final choice of W was based upon several considerations. These were as follows:

- When asked to evaluate the AQoL all-worst health state directly, 76% of respondents rated it as worse than death using the TTO technique.
- In a previous survey, reported earlier, respondents rated W using a visual analogue scale. The average disutility value was 1.00.
- In Tables 5 and 6, nine of the estimates in the columns B had disutility scores greater than 1.0.
- Only two estimates (from states C and D in Table 5) were significantly different from 1.0 and these could have arisen because of the instability of the endogenous estimates.

From these results we concluded that the true value of W was very close to unity. Because of the likelihood of error, the three lowest estimates were discarded and the remainder averaged. This produced a value of 1.04. We have accepted this as the base estimate of the disutility of the AQoL all-worst health state, W. Deflating the unadjusted dimension weights in Table 4 with this estimate results in the AQoL formulas:

$$DU(AQoL)^{LD} = 1.04 - 1.04(1-.613DU_1)(1-.841DU_2)(1-.855DU_3)(1-.931DU_4)(1-.997DU_5)$$

Equation 22

and:

$$U(AQoL)^{LD} = 1 - DU(AQoL) = 1.04(1-.613DU_1)(1-.841DU_2)(1-.855DU_3)(1-.931DU_4)(1-.997DU_5) - 0.04$$

Equation 23

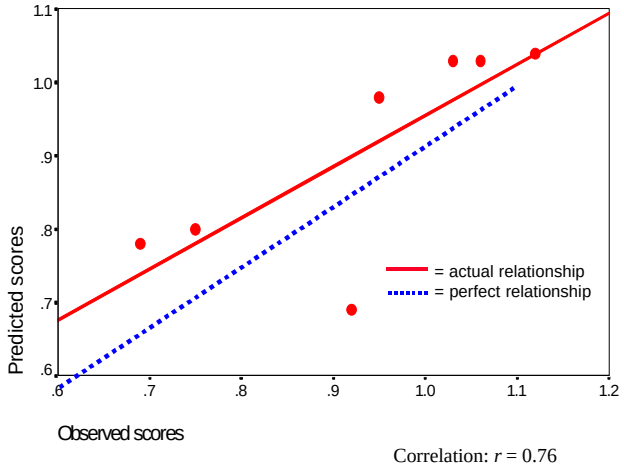
The relationship between the various multi attribute health states that were directly observed during the scaling survey and the values predicted by this final model are shown in Table 7. While these health states were used to help determine the value of the AQoL all-worst, each of the states was measured independently and the goodness of the model fit may, therefore, be taken as a test of the model's predictive power. By this criterion the final model performs well.

Table 7: Observed and predicted multi-attribute health state disutilities, using the final AQoL model

Health state	*AQoL definitions*	*'Observed' Disutility*[1]		
		Observed	*Predicted*	*Obs/Pred.*
I Dimension Worst States				
Dimension 1	(444, 111, 111, 111, 111)	0.64	0.64	1.0
Dimension 2	(111, 444, 111, 111, 111)	0.88	0.88	1.0
Dimension 3	(111, 111, 444, 111, 111)	0.89	0.89	1.0
Dimension 4	(111, 111, 111, 444, 111)	0.97	0.97	1.0
Dimension 5	(111, 111, 111, 111, 444)	1.04	1.04	1.0
II Other composite health states				
A	(444, 444, 111, 111, 111)	0.95	0.98	0.97
B	(111, 111, 444, 444, 111)	1.06	1.03	1.03
C	(333, 111, 333, 111, 432)	0.75	0.80	0.94
D	(111, 333, 111, 313, 432)	0.69	0.78	0.88
E	(141, 411, 114, 114, 114)	1.03	1.03	1.00
F	(114, 141, 141, 141, 141)	0.92	0.69	1.33
W	(444, 444, 444, 444, 444)	1.12	1.04	1.08
Mean of composites				1.03

The five dimension 'all worst' health states are perfectly predicted because of the construction of the AQoL. However, the six 'other' interior points A-F were independently measured and their close correspondence with the predicted utility scores indicates that the model performs well. The relationship between the observed and predicted utilities for these 'other' health states is plotted in Figure 5. The Pearson correlation coefficient between the observed and predicted scores is 0.76.

Figure 5: Observed and predicted (model) disutilities for selected multi-attribute health states



The scoring algorithm of the AQoL incorporates the findings above with respect to the handling of negative utilities and the permitted lower boundary.

5 Discussion

Adjustment of the base model

The estimate of the AQoL all-worst health state and the AQoL scoring system incorporated in the final model was based upon the assumption that estimates of W only differed because of random error. Inspection of the results reported above indicates that this is untrue. Figure 5 shows not only the relationship between the observed and predicted values of W, but also the line of 'perfect' prediction. From this it is clear that lower disutilities are exaggerated by the model and the opposite is true for large disutilities. In part, although not in whole, this relationship is an artifact arising from our selection of a model all-worst value (DU = 1.04) which is less than the directly observed upper value (DU = 1.12).

It was noted earlier that estimates of W would be extremely sensitive to errors in the observed value of utilities. This is particularly obvious from Equation 21 where an error in the observed value, a, has a multiplied impact upon W and in the same direction as the error; that is, a small positive error in a will have a larger percentage and positive impact upon W. The relationships in Figure 5 could, therefore, be explained if there was a systematic upward bias in a, the directly observed disutility, relative to any error incorporated in lower disutility scores.

A likely source of bias which would explain these results is the existence of a 'shock horror' effect amongst respondents when they first contemplate living in very poor health states. The effect of such a systematic bias would be to exaggerate the larger observed DU scores in Figure 5. A shock horror effect would also systematically exaggerate the implied estimate of W as the health state from which W was estimated became worse. For the better health states (C and D; Table 7) the directly observed value, a, would be relatively unaffected. However the disutility predicted from the model, a^1, depends upon the dimension utility scores and the corresponding dimension disutility weights, w_i. These are derived from the dimension all-worst health states and are worse than states C and D, with the exception of Dimension 1. That is, they are likely to incorporate a greater 'shock horror' bias than the direct observations a. The consequence would be that the value of $W = a/a'$ would be biased downwards as the numerator is inflated by less than the denominator as a result of the postulated bias.

In the case of endogenous estimates there would be a similar effect. This may again be seen from Equation 21 from which $dW/da > 0.00$. For better health states a, where estimates of a are relatively undistorted but the dimension weights incorporate an upward bias, the estimated values of W will be biased downwards. However for worse health states where the relative (upward) bias of a is greater, then W will be biased upwards.

A second possible explanation of the systematic relationships in Table 7 and Figure 5 is that the multiplicative model used provides a partial, but not perfect, explanation of the way in which different dimensions of disutility are combined (the assumption of mutual utility independence may be violated) (Feeny, Torrance et al. 1996). As noted by von Winterfeldt and Edwards, and illustrated in the text, the multiplicative model is limited in its flexibility as it permits only one additional degree of freedom when compared with the additive model (von Winterfeldt and Edwards 1986). This degree of freedom is of vital importance for MAU modeling as it permits the disutility of each dimension to reduce the overall HRQoL to a state close to death (Richardson

and Hawthorne 1998). Nevertheless there is no empirical or logical reason why the model fit must be precise. This implies that the first order estimate of utility scores incorporated in the Equations 22 and 23 may need, subsequently, to be adjusted to obtain more accurate, second order, estimates of true utility. This possibility is to be investigated in future research.

6 Conclusions

This paper has considered two issues that have received very little attention in the literature despite their importance for the numerical values produced by MAU instruments. Both concern the treatment of negative values.

An explicit or implicit decision must be made with respect to the treatment of negative values in any generic instrument which evaluates health states close to death. If, as with the AQoL, an instrument contains several dimension each of which are found, empirically, to reduce overall utility to a state close to death then the combination of these dimensions will almost certainly result in at least some respondents to a survey nominating negative values. A decision concerning the treatment of these values can only be avoided by the use of a scaling instrument that does not record negative scores. This is equivalent to a devaluation of the preferences of individuals who regard certain health states as being significantly worse than death. This, of course, does not really avoid the problem but imposes one particular solution; *viz*., to impose the same preference score for a range of true (negative) preferences. If, as most accept, there can be meaningful negative utilities then this 'solution' is arbitrary and hard to justify.

Once negative utilities are permitted two decisions must be made (implicitly or explicitly). The first concerns the lowest negative score which will be accepted as having meaning. The second concerns the utility values on a conventional scale that will be assigned to negative scores, and, particularly, the utility value which will be assigned to the instrument 'all-worst' health state. At present, there is no gold standard criterion — or even accepted guidelines — for the determination of appropriate numerical values. The solution adopted for the scaling of the AQoL and outlined in this paper has been guided by available evidence on people's reaction to worse than death health states as measured by both the TTO and VAS scaling techniques, but ultimately it has necessarily reflected the authors' judgement about the maximum disutility which is consistent with the preservation of a meaningful interval property on the utility scale.

Both of these issues relating to the treatment of negative scores have a quantitatively important impact upon final utility scores and as they deal with the life/death exchange rate (which is the defining characteristic of the QALY) they are issues which are fundamental to the validity of an instrument.

Results from the AQoL scaling study reported here are consistent with the hypothesis of a systematic upward bias in the disutility estimates of poor health states. Alternatively, the data are consistent with the view that the relationship between multi-attribute health states and true utility on a life/death scale is somewhat more complex than implied by the models used to date and that the assumption of mutual utility independence is partially violated. While the multiplicative model is clearly more flexible than the additive model it still imposes a comparatively simple structure upon utilities and a structure which may result in the need to correct predicted utility scores.

Despite these caveats the evidence presented in this paper suggest that the selected AQoL model provides good first order estimates of true utility. The evidence of systematic bias arising from the modeling suggests the need for a second order correction.

References

Dolan P, Gudex C, Kind P*, et al.* (1995). Social tariff for EUROQoL: results from a UK general population survey. Discussion Paper 138. York: Centre for Health Economics, University of York.

Feeny D, Torrance G and Furlong W (1996). Health utilities index. In B. Spilker. Quality of Life and Pharmacoeconomics in Clinical Trials. Philadelpha: Lippincott-Raven Publishers.

Furlong W, Feeny D, Torrance G*, et al.* (1998). Multiplicative Multi-attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report. 98-11. Hamilton: McMaster University, Centre for Health Economics and Policy Analysis.

Hawthorne G, Richardson J and Day N (2000). Using the Assessment of Quality of Life (AQoL) Instrument. Technical Report 12. Melbourne: Centre for Health Program Evaluation.

Hawthorne G, Richardson J, Day N*, et al.* (2001). Utility Weights for the 'Assessment of Quality of Life' (AQoL) Instrument. Melbourne: Centre for Health Program Evaluation.

Hawthorne G, Richardson J, Day N*, et al.* (2000). Construction and Utility Scaling of the Assessment of Quality of Life (AQoL) Instrument. Working Paper 101. Melbourne: Centre for Health Program Evaluation.

Hawthorne G, Richardson J and Osborne R (1999). The Assessment of Quality of Life (AQoL) Instrument: a psychometric measure of health related quality of life. *Quality of Life Research.* 8: 209-224.

Hawthorne G, Richardson J, Osborne R*, et al.* (1997). The Australian quality of life (AQoL) instrument: initial validation. Working Paper 76. Melbourne: Centre for Health Program Evaluation.

Merbitz C, Morris J and Grip J (1989). Ordinal scales and foundations of misinference. *Archives of Physical and Medical Rehabilitation.* 70: 308-312.

Nord E, Richardson J and Macarounas-Kirchmann K (1993). Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. *International Journal of Technology Assessment in Health Care.* 9: 463-78.

Richardson J (1994). Cost utility analysis: what should be measured? *Social Science and Medicine.* 39: 7-21.

Richardson J and Hawthorne G (1998). Difficulty with life and death: methodological issues and results from the utility scaling of the 'Assessment of Quality of Life' (AQoL) instrument. Economics and Health: 1988 Proceedings of the Twentieth Australian Conference of Health Economists, Sydney: The University of Sydney, School of Health Services Management.

Rosser R and Kind P (1978). A scale of valuations of states of illness: is there a social consensus. *International Journal of Epidemiology.* 7: 4-15.

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

Torrance G, Boyle M and Horwood S (1982). Application of multi-attribute theory to measure social preferences for health states. *Operations Research.* 30: 1043-1069.

Torrance G, Furlong W, Feeny D*, et al.* (1995). Multi-attribute preference functions: health utilities index. *PharmacoEconomics.* 7: 503-520.

von Winterfeldt D and Edwards W (1986). Decision analysis and behavioural research. Cambridge: Cambridge University Press.

Williams A (1995a). The measurement and valuation of health: a chronicle. 136. York: Centre for Health Economics.

Williams A (1995b). The measurement and valuation of health: final report on the modelling of valuation tariffs. York: MVH Group, Centre for Health Economics, University of York.

Appendix 1:

Assessment of Quality of Life (AQoL) Instrument

INSTRUCTIONS:

Please circle the alternative that best describes you *during the last week*.

1 Concerning my use of prescribed medicines:
- A. I do not or rarely use any medicines at all.
- B. I use one or two medicinal drugs regularly.
- C. I need to use three or four medicinal drugs regularly.
- D. I use five or more medicinal drugs regularly.

2 To what extent do I rely on medicines or a medical aid? (NOT glasses or a hearing aid.) *(For example: walking frame, wheelchair, prosthesis etc.)*
- A. I do not use any medicines and/or medical aids.
- B. I occasionally use medicines and/or medical aids.
- C. I regularly use medicines and/or medical aids.
- D. I have to constantly take medicines or use a medical aid.

3 Do I need regular medical treatment from a doctor or other health professional?
- A. I do not need regular medical treatment.
- B. Although I have some regular medical treatment, I am not dependent on this.
- C. I am dependent on having regular medical treatment.
- D. My life is dependent upon regular medical treatment.

4 Do I need any help looking after myself?
- A. I need no help at all.
- B. Occasionally I need some help with personal care tasks.
- C. I need help with the more difficult personal care tasks.
- D. I need daily help with most or all personal care tasks.

5 When doing household tasks: *(For example, preparing food, gardening, using the video recorder, radio, telephone or washing the car)*
- A. I need no help at all.
- B. Occasionally I need some help with household tasks.
- C. I need help with the more difficult household tasks.
- D. I need daily help with most or all household tasks.

6 Thinking about how easily I can get around my home and community:
- A. I get around my home and community by myself without any difficulty.
- B. I find it difficult to get around my home and community by myself.
- C. I cannot get around the community by myself, but I can get around my home with some difficulty.
- D. I cannot get around either the community or my home by myself.

7 Because of my health, my relationships *(for example: with my friends, partner or parents)* generally:
- A. Are very close and warm.
- B. Are sometimes close and warm.
- C. Are seldom close and warm.
- D. I have no close and warm relationships.

8 Thinking about my relationship with other people:

A. I have plenty of friends, and am never lonely.
B. Although I have friends, I am occasionally lonely.
C. I have some friends, but am often lonely for company.
D. I am socially isolated and feel lonely.

9 Thinking about my health and my relationship with my family:
A. My role in the family is unaffected by my health.
B. There are some parts of my family role I cannot carry out.
C. There are many parts of my family role I cannot carry out.
D. I cannot carry out any part of my family role.

10 Thinking about my vision, including when using my glasses or contact lenses if needed:
A. I see normally.
B. I have some difficulty focusing on things, or I do not see them sharply.
For example: small print, a newspaper, or seeing objects in the distance.
C. I have a lot of difficulty seeing things. My vision is blurred.
For example: I can see just enough to get by with.
D. I only see general shapes, or am blind. *For example: I need a guide to move around.*

11 Thinking about my hearing, including using my hearing aid if needed:
A. I hear normally.
B. I have some difficulty hearing or I do not hear clearly.
For example: I ask people to speak up, or turn up the TV or radio volume.
C. I have difficulty hearing things clearly. *For example: Often I do not understand what is said. I usually do not take part in conversations because I cannot hear what is said.*
D. I hear very little indeed. *For example: I cannot fully understand loud voices speaking directly to me.*

12 When I communicate with others: *(For example: by talking, listening, writing or signing)*
A. I have no trouble speaking to them or understanding what they are saying.
B. I have some difficulty being understood by people who do not know me. I have no trouble understanding what others are saying to me.
C. I am only understood by people who know me well. I have great trouble understanding what others are saying to me.
D. I cannot adequately communicate with others.

13 If I think about how I sleep:
A. I am able to sleep without difficulty most of the time.
B. My sleep is interrupted some of the time, but I am usually able to go back to sleep without difficulty.
C. My sleep is interrupted most nights, but I am usually able to go back to sleep without difficulty.
D. I sleep in short bursts only. I am awake most of the night.

14 Thinking about how I generally feel:
A. I do not feel anxious, worried or depressed.
B. I am slightly anxious, worried or depressed.
C. I feel moderately anxious, worried or depressed.
D. I am extremely anxious, worried or depressed.

15 How much pain or discomfort do I experience?
A. None at all.
B. I have moderate pain.
C. I suffer from severe pain.
D. I suffer unbearable pain.