MONASH University
Business and Economics

Centre for Health Economics

Working Paper 143

The Assessment of Quality of Life (AQoL) II Instrument

The Effect of Deliberation and Alternative Utility Weights in a Multi-Attribute Utility Instrument

Dr Stuart Peacock
Senior Lecturer, Health Economics Unit, Monash University

Professor Jeff Richardson
Director, Health Economics Unit, Monash University

Mr Neil Atherton Day
Principal Research Fellow, Centre for Program Evaluation, The University of Melbourne

Mr Angelo Iezzi
Research Fellow, Health Economics Unit, Monash University

March, 2004
ISSN 1448 - 6822
ISBN 1 876662 64 6

PREFACE

Economic evaluation of health and health care related activities must quantify the *importance* of the quality of life of the outcome. This is done through the use of the Quality Adjusted Life Year (QALY) as a unit of outcome. As QALYs are defined as the product of life years and an index of the quality of life, this latter quantity requires measurement. This task is carried out by measuring the *strength of preference* for a health state relative to full health and death.

A number of utility instruments presently exist but the utility scores produced by them differ very significantly (Hawthorne et al 2001). The AQoL project was a response to the implied challenge. It was undertaken in an attempt to improve the methodology of MAU instrument construction and, hopefully to produce utility scores which had greater reliability and validity.
Innovations include the following:

(i) The descriptive system

- use of the correct psychometric procedures for instrument construction;
- a description based upon 'handicap'—problems in a social context—as distinct from a 'within the skin' descriptive system;
- a multi level descriptive system which permitted redundancy—double counting—within dimensions in order to achieve instrument sensitivity but structural independence between the dimensions;
- an increase in the sensitivity of the descriptive system in the region of full health and specifically a system which permits the evaluation of health promotional activities.

(ii) Tariffs (Scaling/Calibration)

- The creation of at least two scaling systems based upon the time trade-off (as with AQoL 1) and the person trade-off (PTO) scaling methodologies. The appropriate choice of scaling instrument has not been determined in the literature;
- A reworking of the utility scores employing techniques to eliminate one possible source of bias in previous methodologies (including AQoL 1), viz a 'focusing effect';
- The testing and use of 'deliberative weights' which permit and encourage the contemplation of the health states over a significant period time;
- The use of a more flexible two stage modelling methodology to combine disaggregated dimension scores into an overall utility score for a multi attribute health state.

Results from the AQoL 2 project are published in four reports. These are:

(i) Conceptualising the Assessment of Quality of Life Instrument Mark 2 (AQoL 2), Methodological Innovations and the Development of the AQoL Descriptive System;

(ii) The Assessment of Quality of Life (AQoL) II Instrument: Derivation of the scaling weights using a multiplicative model and econometric second stage correction;

(iii) The Assessment of Quality of Life (AQoL) II Instrument: The effective deliberation and alternative utility weights in a multi attribute utility instrument;

(iv) Overview of the Assessment of Quality of Life Mark 2 Project.

TABLE OF CONTENTS

1 Introduction....1

Issues in Multi Attribute Utility (MAU) Modelling....1

2. Deliberation - Methods....3

3. Deliberation - Results....7

4. The Person Trade-Off (PTO)....12

References....17

Appendix 1 AQoL Interviewer Instructions....18

Appendix 2 Visual Aid for PTO Questions....25

List of Tables

Table 1	Test for Self-selection and over-sampling: Interview I control vs deliberative arms	8
Table 2	Test of deliberation: pooled data	9
Table 3	Test of deliberation: control arm	9
Table 4	Test of deliberation: deliberation arm	9
Table 5	Test of deliberation: pooled data	10
Table 6	TTO Disutilities – pooled data	10
Table 7	TTO Disutilities – control arm	11
Table 8	TTO Disutilities – deliberative arm	11
Table 9	Scaling Survey: Respondents and Response Rates	13
Table 10	PTO-TTO Transformatio Regression Results	14
Table 11	AQoL Population Norms Disutility	15
Table 12	Population norms AQoL Disutility, PTO Disvalue	15
Table 13	PTO Population Norms (Dis Value)	15
Table 14	Standardised PTO Values	15

List of Figures

Figure 1	AQoL II Survey Design	4
Figure 2	Survey Recruitment Matrix – Numbers of Respondents	4
Figure 3	AQoL Model – TTO Disutilities	14
Figure 4	AQoL Model – PTO Disutilities	14
Figure 5	Average of utilities versus utility of the average	16
Figure 6	Transformation of TTO into PTO values	16

The Assessment of Quality of Life (AQoL) II Instrument
The Effect of Deliberation and Alternative Utility Weights in a Multi-Attribute Utility Instrument

1 Introduction

Issues in Multi Attribute Utility (MAU) Modelling

MAU instruments have been in use for 30 years. Despite this there remain a very large number of issues with respect to their reliability and their validity, that is the question of whether or not they measure what they purport to measure which, for an MAU instrument, is utility. An incomplete list of these problems is given below.

Descriptive system

The construct 'health' may be described several ways. Using old WHO terminology, a descriptive system may be based upon handicap, impairment and disability. The latter two are often described as 'within the skin' descriptions as they attempt to describe health with reference to only the functioning of a person's body and mind. Handicap is the description of a person's health with reference to their social context. AQoL adopted this latter framework and thereby embodied the hypothesis that handicap, not impairment or disability is the primary determinant of wellbeing.

The descriptive system—set of questions and response categories—should achieve preference and structural independence. The former refers to the interdependence of preferences. Results will be invalidated if the preference for one dimension of health depends significantly upon the level of health measured along another dimension. The latter requirement is that dimensions do not 'overlap'; that is, an element of the description should be captured by only one item. This requirement translates into the standard psychometric requirement that dimensions be orthogonal.

Finally, a descriptive system should be sensitive; that is, it should be capable of detecting changes in a health state sufficiently small that the objective of the instrument is achieved. However, a trade-off exists between this and the previous requirement. As health states are described in greater detail there is an increasing likelihood of overlap.

Scaling

Scaling—calibrating the descriptive system—also encounters conceptual and practical difficulties. First, because the number of health states is potentially vast it is not possible to measure every combination of dimensions separately. It is for this reason that MAU instruments have adopted the decision analytic solution of modelling health, ie decomposing the construct into dimensions scaling these dimensions and then recombining them into a multi attribute—composite—health state description. The recombination of dimensions may use additive or multiplicative models and both of these may be based upon either mathematical averaging or statistical analysis. The

AQoL project has, for the first time, adopted a two stage strategy in which a multiplicative model is used first and then correction sought in a second stage econometric analysis.

A further set of issues arise from the alternative perspectives which may be adopted in scaling. Either patients or a cross section of the public may be asked to provide their preferences. They may be asked to imagine that they personally are in the health state described. Alternatively they may be asked to provide an abstract judgement on the social benefits of curing abstract patients from the health state described.

Finally the universal practice to date has been to seek 'spontaneous' responses—the research cost of surveying is sufficiently high that there is little time for respondents to ponder and assess the consequences of being in the health state which is described to them.

AQoL 2 has been used as a vehicle for experimentation with some of these issues. In particular, and as noted above, it has pioneered a two stage modelling procedure. First, it has adopted a multi level descriptive system. Secondly, it has adopted a 2 stage scaling procedure. These are described in two earlier reports (Richardson 2003). The present paper reports the results of two such innovations. These are, first, the result of a structured method for eliciting deliberation from respondents. The second is the use of two scaling techniques, viz, the time trade-off and the person trade-off. To date all of the MAU instruments have adopted a personal perspective and asked respondents to imagine they were in a health state. The strength of preference for avoiding this state is then measured using a rating scale (15D; the standard gamble (HUI III) or the time trade-off technique (AQoL 1, SF36) Brazier weights (EQ5D). The AQoL 2 is unique in collecting data on the person trade-off (PTO).

Deliberation

As noted none of the MAU instruments in the literature have attempted to employ 'deliberative' weights, ie utility scores derived from interview respondents who have considered the questions at length. More generally this issue has been almost totally ignored in the utility literature.

2. Deliberation - Methods

A key feature of the AQOL II has been the development of spontaneous and deliberative utility weights. Deliberation refers to either (Fearon 1998):

1. A particular sort of discussion – one that involves the careful and serious weighing of reasons for and against some proposition.
2. Or an interior process by which an individual weighs reasons for and against courses of action.

To date, all QALY (Quality Adjusted Life Year) scores have been based on MAU instruments which have used a single interview format to elicit spontaneous utility weights from a population sample. However, the universal practice of obtaining utility scores from a single interview has been questioned in two significant studies.

The WHO DALY (Disability Adjusted Life Year) study (Murray and Lopez 1996) replaced a single interview strategy with a two stage procedure. Initial surveys were administered using PTO (Person Trade-Off) and TTO (Time Trade-Off) techniques. These were followed by focus groups of expert respondents discussing inconsistencies and conflicts from earlier responses, producing deliberative utility scores. Spontaneous and deliberative scores were found to be virtually uncorrelated.

Sheill et al (2000) obtained valuations for two health states over three separate interviews held within a 1 to 8 week period using a convenience sample of Medical Faculty staff. 36% of participants showed evidence of reflection i.e. the first interview prompted them to think about their valuations and to change their answers at the second interview. 24% of participants showed evidence of reflection at the third interview. Only 40% of participants had stable valuations over time, suggesting spontaneous valuations may be unreliable.

The findings raise questions about the validity of QALY scores based on spontaneous utility weights and the degree of correspondence between repeat administration of valuation questions.

AQOL II Deliberative Survey Design

Utility weights for the AQOL II were elicited from respondents over two face-to-face interviews. The study adopted a novel survey design to test the reliability of TTO valuations (test-retest) and to test for differences between spontaneous and deliberative TTO valuations. The survey design is shown in figure 1.

The population sample was randomly drawn from the Victorian population using the White Pages. Potential respondents were contacted by mail, and those agreeing to participate were stratified into one of five SEIFA groups based on postcode of residence. The study sought to recruit approximately 400 participants, to obtain valuations from 360 respondents (allows for 10% drop out/invalid responses).

All respondents were asked to complete the AQOL I and AQOL II questionnaires, a baseline socio-economic and demographic questionnaire. All respondents completed 11 TTO questions at interview I: 6 TTOs for AQOL dimension worst health states, 1 TTO for the AQOL all-worst health state, 3 TTOs on AQOL multi-attribute health states, and 1 TTO on their own, current, health state. 6 sorts of TTO questions were used. Each sort contained the same dimension worst TTOs, the AQOL all-worst TTO and the own health TTO, but the order of the dimension worst TTOs was

varied to remove any bias from question ordering effects. Each sort contained a different set of 3 multi-attribute health state TTOs (making a total of 18 multi-attribute health state questions used in the study) which were used for econometric modelling. Respondents were assigned to one of the six sorts prior to interview I.

At the end of interview I, respondents were randomly assigned to either the control or the deliberative arm of the study. This led to the survey recruitment matrix shown in figure 2.

Figure 1 AQOL II Survey Design

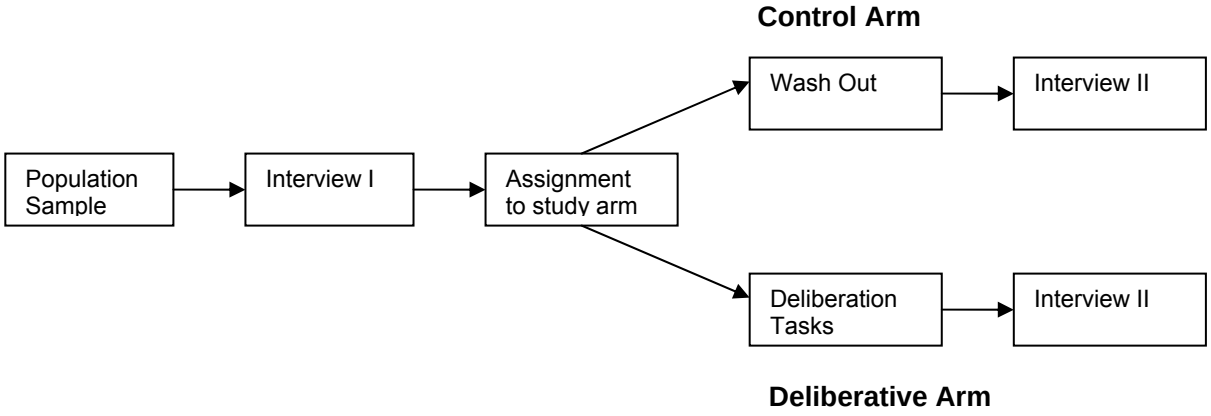


Figure 2 Survey Recruitment Matrix – Numbers of Respondents
(C = control arm, D = deliberative arm)

	Sort 1		Sort 2		Sort 3		Sort 4		Sort 5		Sort 6		
	C	**D**	**C**	**D**	**C**	**D**	**C**	**D**	**C**	**D**	**C**	**D**	**TOTAL**
SEIFA Group1	6	6	6	6	6	6	6	6	6	6	6	6	72
SEIFA Group2	6	6	6	6	6	6	6	6	6	6	6	6	72
SEIFA Group3	6	6	6	6	6	6	6	6	6	6	6	6	72
SEIFA Group4	6	6	6	6	6	6	6	6	6	6	6	6	72
SEIFA Group5	6	6	6	6	6	6	6	6	6	6	6	6	72
TOTAL	30	30	30	30	30	30	30	30	30	30	30	30	360

Respondents assigned to the control arm were simply told to come back for Interview II in 2-3 weeks time. The 2-3 week period was therefore used as a wash-out period, during which respondents should forget their responses from interview I. This allowed the testing of test-retest reliability of valuations made at interview I compared to valuations made at interview II.

Respondents assigned to the deliberative arm were given a set of “deliberation tasks” to complete in the 2-3 week period before interview II. These tasks were designed to stimulate reflection and further consideration of valuations before interview II. This allowed testing of the effects of

deliberation on valuations made at interview I compared to valuations made at interview II. The deliberation tasks asked respondents in the deliberative arm to:

- Complete the AQOL II questionnaire for the worst health state they had ever experienced.
- Complete a TTO on the worst health they had ever experienced.
- Complete a TTO on the AQOL all-worst health state.
- Complete a TTO on one of the AQOL dimension worst health states.

To do this, deliberative respondents were provided with a blank AQOL II questionnaire and TTO sheets for self completion at home. Critically, deliberative respondents were asked to discuss the questions, their answers, and their reasons for their valuations (e.g. what particular aspect of a health state made it so bad) with a family member or friend. Respondents were asked to speak with the person they might discuss health related problems with in real life, e.g. a spouse, close relative or close friend etc. These tasks were based on the rationale that:

- Imagining the worst health state they had ever experienced should encourage greater critical reflection on different aspects of health, and recollection of personal experiences should make the descriptions of AQOL health states more meaningful to the individual.
- Discussion with a family member or friend who they would turn to with real life health related problems should also encourage greater critical reflection, should broaden issues considered in arriving at valuations, and should clarify thinking on different aspects of health.

It was also recognised that, in real life, some individuals may choose not to discuss health problems with a family member or friend (e.g. due to social isolation, or because they prefer their own counsel). It was important that these people were not excluded from the deliberative tasks, because deliberation was designed to encourage the type of reflection that would occur in real life situations. Therefore, at the end of interview I deliberative respondents were asked if they would choose to discuss health problems with anyone in real life, and if they would not they were asked to complete the deliberation tasks just as if they were facing a real life trade-off decision on their own.

At then end of interview I, interviewers asked all respondents assigned to the deliberative arm whether they would be willing to complete the deliberation tasks. Those who were not willing were assigned to the control arm. Those who had found interview I extremely emotionally or cognitively demanding were also assigned to the control arm. Respondents in the deliberative arm spent an additional 10 minutes at the end of interview I having the deliberation tasks explained to them by the interviewer. Respondents were asked to bring their completed deliberation materials with them to interview II.

At interview II all respondents (control and deliberative) completed the same 11 TTOs they had answered in interview I. Those in the deliberative arm also had the opportunity to discuss the deliberation tasks and any changes to their responses with the interviewer. Note, all respondents also completed 10 PTOs at the end of interview II (see later working paper).
The AQOL II survey design therefore provides a unique set of tests for the reliability of, and the effects of deliberation on, TTO valuations:

- Comparison between TTOs from interview I in the control arm and TTOs from interview I in the deliberative arm provides a check for any systematic differences from any self selection between survey arms.

- Comparison between TTOs from interview I and II in the control arm provides a “pure” test of TTO test-retest reliability (with wash-out between interviews).
- Comparison between TTOs from interview I and II in the deliberative arm provides a joint test of TTO test-retest reliability and of deliberation on TTO valuations.
- Comparison between TTOs from interview II in the control arm and TTOs from interview II in the deliberative arm provides a “pure” test of deliberation on TTO valuations (net of any test-retest reliability effects).

3. Deliberation - Results

Of the 366 individuals who provided valid TTO responses at both interviews I and II, 143 successfully computed the deliberative tasks. The remaining 223 participants were included in the control arm. The over-sampling evident in the control arm of the study was a result of the three exclusion criteria listed above:

- participant unwilling to undertake deliberative tasks
- participant found interview I very emotionally or cognitively demanding
- participant agreed to do deliberative tasks, but failed to complete them.

However, over-sampling in the control arm was not a problem, as survey design specifically allowed for testing of self-selection effects. Moreover, over-sampling of control is common in most epidemiological study designs.

Analysis of test-retest reliability and the effects of deliberation is based on the dimension-worst and AQoL all worst TTO valuations obtained at interview I and II for pooled data (both arms of the study) and for control and deliberative arms separately. Descriptive statistics for pooled, control and deliberative participants are provided below.

(i) Test for effects of self-selection and over-sampling

Comparison of the TTO valuations from Interview I in the control and deliberative arms provides a check for any systematic differences that nay have arisen due to self-selection and over sampling in the control arm.

The most commonly used method for testing differences between survey or trial arms is to test for differences in mean values using a pooled t test (also called a student t test or independent simple t test).

A critical assumption of the pooled t test is that the variance of TTO valuations in both arms of the survey are equal. Accordingly, Levine's test for equality of variances was performance on the 6 dimension-worst TTOs and the AQoL all-worst TTOs across the control and deliberative arms. Variances were found to be equal for dimension worst C1, C2 and C6 and for the AQoL all worst. A pooled variance t test for differences in mean TTO valuations was performed on these data. Variances were not found to be equal for dimension worsts C3, C4 and C5, and separate variance t tests were performed on these data.

A further, less critical, assumption of the pooled t test is that the distribution s of the TTO valuations in each arm of the survey are normal. To allow for the potential for non-normality in these distributions, a Pearson's chi-squared test was also performed. This test is robust to many forms of data and distributions.

Table 1 Test for self-selection and over-sampling: Interview I control vs deliberative arms

		Pooled t test		Pearson's chi-squared	
		t	sig(2 tailed)	t	sig(2 tailed)
Dimension worst	TTO C1	1.066	0.287	42.12	0.160
	TTO C2	-2.763	0.006	26.11	0.759
	TTO	-0.506	0.613	42.64	0.080
C3*					
	TTO	0.150	0.881	36.77	0.219
C4*					
	TTO	-1.971	0.050	31.99	0.567
C5*					
	TTO C6	-1.011	0.313	50.37	0.126
AQoL all worst TTO DAQ		-0.838	0.403	47.27	0.171

* t test based on separate variance t test

Results show that pooled t tests indicate possible systematic differences between control and deliberative arms in valuations for C2 and C5 dimension worsts. However, the more robust Chi-squared test do not show any systematic differences due to self-selection and over-sampling in the control arm.

(ii) Test-retest reliability

Comparison of the TTO valuations between interviews I and II provides a test of the test-retest reliability of repeat administration of TTO questions. Results are presented in 3 groups: pooled data (n = 366) control arm (n = 233) and deliberative arm (n = 143). Pooled analysis provides an overall assessment of reliability, analysis of the control arm allows for a 'pure' test of reliability after a wash-out period, and analysis of the deliberative arm allows for testing the effects of deliberative tasks in conjunction with reliability.

The most commonly used method for analysis of dependent samples (matched or paired data) is to use a paired sample test for the mean difference. A significant result ($p < 0.05$) implies that the mean TTO values differed across interview I and II, ie TTO valuations may not be reliable under repeat administration.

An alternative non-parametric test, robust to many forms of non-normality, is the Wilcoxon Signed Ranks test. This methods tests whether repeat administration has an effect on TTO valuations based on observed changes in respondent's valuations over time. The test statistic is based on the standard and normal ($\not{Z}$) distribution, and a p value < 0.05 implies TTO valuations change after repeat administration.

Shiell et al (2000) have proposed a further test for test-retest reliability. They use the intra class correlation coefficient as a recommended indicator of agreement between administrations.

Results for pooled data and each survey arm are presented below. The three methods for assessing test-retest reliability produce broadly similar results: that TTO valuations may vary over repeat administrations. Pooled data indicate that TTO valuations for three of the dimension worsts and the AQoL all worst differ between interviews I and II, but 5 of the 7 intra class correlations indicate good agreement (>0.5) under standard test-retest criteria.

Results for the control arm—the 'pure' test of test-retest reliability—indicate changes in two of the dimension worsts and also in the AQoL all worst. Results for the deliberative arm also suggest this pattern.

Table 2 Test of deliberation: pooled data

(a) Pooled Data (n = 366)		Paired t test		Wilcoxon Signed Ranks Test		Intra class correlation coefficient
		t	sig(2 tailed)	Z	sig(2 tailed)	
Dimension worst	TTO C1	3.77	0.00	-3.49	0.00	0.53
	TTO C2	2.88	0.00	-2.74	0.01	0.54
	TTO C3	1.70	0.09	-1.69	0.09	0.54
	TTO C4	0.13	0.90	-0.37	0.71	0.47
	TTO C5	-2.48	0.01	-2.39	0.02	0.48
	TTO C6	1.97	0.05	-1.81	0.07	0.56
AQoL all worst	TTO DAQ	2.68	0.01	-2.64	0.01	0.51

Table 3 Test of deliberation: control arm

(b) Control Arm (n = 233)		Paired t test		Wilcoxon Signed Ranks Test		Intra class correlation coefficient
		t	sig(2 tailed)	Z	sig(2 tailed)	
Dimension worst	TTO C1	2.54	0.01	-2.13	0.03	0.53
	TTO C2	3.83	0.00	-3.68	0.00	0.56
	TTO C3	1.44	0.15	-1.07	0.28	0.56
	TTO C4	0.05	0.96	-0.44	0.66	0.47
	TTO C5	-1.40	0.16	-1.28	0.20	0.50
	TTO C6	2.03	0.04	-1.60	0.11	0.58
AQoL all worst	TTO DAQ	1.46	0.15	-2.30	0.02	0.41

Table 4 Test of deliberation: deliberation arm

(c) Deliberative Arm (n = 143)		Paired t test		Wilcoxon Signed Ranks Test		Intra class correlation coefficient
		t	sig(2 tailed)	Z	sig(2 tailed)	
Dimension worst	TTO C1	2.86	0.01	-2.96	0.00	0.52
	TTO C2	-0.19	0.85	-0.21	0.83	0.51
	TTO C3	0.90	0.37	-1.30	0.20	0.50
	TTO C4	0.16	0.88	-0.03	0.98	0.47
	TTO C5	-2.24	0.03	-2.26	0.02	0.45
	TTO C6	0.62	0.54	-0.81	0.42	0.55
AQoL all worst	TTO DAQ	2.49	0.01	-1.60	0.11	0.61

Results therefore demonstrate some changes in TTO valuations over repeat administrations. There are at least two potential reasons for this.

(i) the instrument may be unreliable
(ii) participants valuations may change through a process of preference construction/values clarification.

The second possibility represents a form of deliberation—internal reflection in particular—which cannot be controlled for experimentally. It is therefore difficult to conclude which of the two potential causes is more likely to be the underlying reasons for changes in valuations.

Tests for effects of deliberation

Comparison of the TTO valuations from Interview II in the control and deliberative arms provide a test of the effects of deliberation. Testing methods follows those used in testing for self-selection effects, by using pooled + tests and Pearson's chi-squared tests. Results are shown below.

Table 5 Test of deliberation: pooled data

	Pooled T test		Pearson's Chi-squared	
	t	Sig(2 tailed)	Chi-Sq	Sig(2 tailed)
Dimension-worst TTO C1	0.604	0.546	29.10	0.707
TTO C2	-0.555	0.579	41.86	0.092
TTO C3	-0.244	0.807	40.85	0.111
TTO C4	0.069	0.945	26.48	0.547
TTO C5	-1.127	0.260	31.79	0.787
TTO C6	-0.326	0.745	33.30	0.764
All worst TTO DAQ*	-1.623	0.106	42.70	0.398

* t test based on separate variance t test

Results are consistent across parametric and non-parametric tests. Deliberation tasks did not affect TTO valuations. However, this does not rule out the potential for preference construction/values clarification occurring in both survey arms. It does, however, suggest that any preference construction is occurring independently of additional stimuli.

Table 6 TTO Disutilities - Pooled Data

Interview 1				Interview 2			
Variable	Obs	Mean	Std Dev	Variable	Obs	Mean	Std Dev
du(1c1)	366	.53	.37	du(2c1)	366	.47	.32
du(1c2)	366	.49	.37	du(2c2)	366	.44	.32
du(1c3)	366	.50	.35	du(2c3)	366	.47	.32
du(1c4)	366	.34	.29	du(2c4)	366	.34	.25
du(1c5)	366	.53	.39	du(2c5)	366	.59	.37
du(1c6)	366	1.15	.19	du(2c6)	365	.63	.38
du(1daq)	366	1.15	.19	du(2daq)	365	1.12	.21

Table 7 TTO Disutilities - Control Arm

Interview 1				Interview 2			
Variable	**Obs**	**Mean**	**Std Dev**	**Variable**	**Obs**	**Mean**	**Std Dev**
du(1c1)	233	.52	.37	du(2c1)	223	.46	.33
du(1c2)	233	.54	.39	du(2c2)	223	.45	.33
du(1c3)	223	.51	.38	du(2c3)	223	.48	.33
du(1c4)	223	.34	.30	du(2c4)	223	.34	.25
du(1c5)	223	.57	.40	du(2c5)	223	.60	.37
du(1c6)	223	.69	.42	du(2c6)	223	.64	.38
du(1daq)	223	1.16	.19	du(2daq)	223	1.14	.18

Table 8 TTO Disutilities - Deliberative Arm

Interview 1				Interview 2			
Variable	**Obs**	**Mean**	**Std Dev**	**Variable**	**Obs**	**Mean**	**Std Dev**
du(1c1)	143	.56	.38	du(2c1)	143	.48	.31
du(1c2)	143	.42	.34	du(2c2)	143	.43	.31
du(1c3)	143	.49	.30	du(2c3)	143	.47	.29
du(1c4)	143	.35	.26	du(2c4)	143	.34	.25
du(1c5)	143	.49	.35	du(2c5)	143	.56	.37
du(1c6)	143	.64	.41	du(2c6)	142	.62	.37
du(1daq)	143	1.14	.20	du(2daq)	142	1.10	.24

4. The Person Trade-Off (PTO)

The person trade-off was one of the earliest techniques used to compare health states. Initially, it was referred to as the 'equivalence technique' (Torrance 1986). After its initial use, it lost popularity as economists argued that utility was best measured by either the standard gamble or the time trade-off instruments (Richardson 1994). In particular, there was a long debate between advocates of the QALY (TTO based 1 year scenario) and the Healthy Life Year Equivalent (HYE); (Rest of Life, Standard Gamble based). Interest in the PTO was revived primarily as a result of the work by Nord (see in particular Nord 1999). A series of studies were carried out in Australia using the technique to measure the preference for different distributions of health benefits (see Nord and Richardson 1995a; 1995b; 1995c; Richardson and Nord 1997).

Partly because of this work the PTO was adopted by the WHO as the technique for measuring the burden of disease and the subsequent Disability Adjusted Life Years (DALYs) reported in the Global Burden of Disease for every disease and for every country were based upon PTO procedures.

The choice of scaling instrument—PTO or TTO—is not primarily a technical issue. Rather, the choice should be based upon the purpose of the measurement and the social values underlying this purpose. Both techniques may be used to measure the burden of disease, arising from the loss of quality of life in the context of an economic evaluation. In both contexts, the choice depends upon the perspective which is desired. The TTO asks respondents to imagine that they are personally in a health state. The PTO asks respondents to imagine that they are a social decision maker and, by implication, not personally affected. That is, the two instruments incorporate a personal and social perspective respectively. As either instrument could be legitimately used in an economic (or other) evaluation the AQoL 2 has provided two sets of tariffs, one for each of the two scaling instruments.

Person Trade-off Results

Results from the 'fieldwork' are also presented in Working Paper 142. They are summarised in Table 9. As described, an initial letter was sent to selected households and was followed up by telephone. In total 2,244 letters were sent from which 1,319 successful telephone contacts were achieved. The principle reasons for non-contact were first, 261 cases where names on the letter did not correspond with the person contacted and 381 cases where no answer was received. Following successful contact 153 cases were dropped because of language difficulties and 136 because of age or ill health. Six hundred 'not interested' cases were recorded resulting in an overall response rate of 41.7 percent of possible respondents. Individuals were asked, in the first instance, to participate in two sets of interviews. As described these were conducted either at the individual's home or, for the vast majority, interviews were carried out at a common location close to their residence. Due to the volume of data collected the two interviews were followed up with a postal survey which was used to obtain rating scale date on item responses. In total 163 questionnaires were returned representing a response rate of 40 percent.

Table 9 Scaling Surveys: Respondents and Response Rates

		TTO-PTO Interview	Mail Questionnaire
Sample		2244	411
Successful contact		1319	
Possible respondents		1030	
	Respondents	430	180
	Response Rate	41.7%	43.8%
Respondents:			
	Sex % male	35%	31%
	Age % age <25	3%	3%
	Age % age ≥60	22%	33%
Education			
	A Primary %	3%	4%
	B Secondary %	47%	50%
	C Tertiary %	50%	46%

As with previous AQoL surveys there was a disproportionate number of women and highly educated respondents. By design, there were no respondents under the age of 25 years. Interviewers did not find the difficulty with elderly respondents that has been reported by some research teams. This may reflect the care with which the interviews were constructed and the particular visual aids that were employed and, particularly, the use of a visual image to represent the mix of good health and poor health states.

Results from the regression analysis are presented in Table 10. The best fitting equation (equation 2) is

$$PTO = TTO^{0.68 + 0.19Slope1 + 0.23\ Slope2 + 0.45\ Slope3}$$

where 'slope 1-3 are dummy variables for observations in the range (0.25 - 0.5), (0.5 - 0.75), (0.75 - 1.0)

The R^2 (fit) of this equation is only marginally superior to equation (1) and, consequently, equation (1) was used to generate PTO values. Equation 1 is:

$$PTO = TTO^{0.72}$$

Population Norms

Population norms for both AQoL (TTO) and AQoL (PTO) are repeated in Tables 11 to 13.

Table 10 **PTO-TTO Transformation Regression Results**

	Dependent: Ln PTO	
Independent variable	**1**	**2**
Ln AQoL	0.72	0.68
AQoL - Slope 1		0.19
AQoL - slope 2		0.23
AQoL - Slope 3		0.45
R^2	0.54	0.55

Figure 3 **AQoL Model - TTO Disutilities**

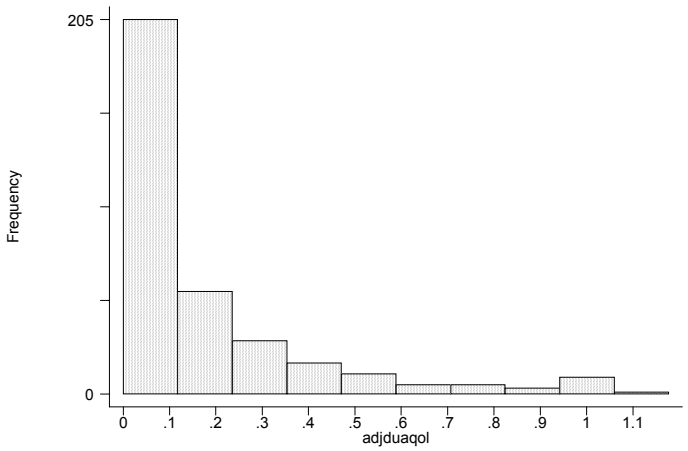


Figure 4 **AQoL Model - PTO Disutilities**

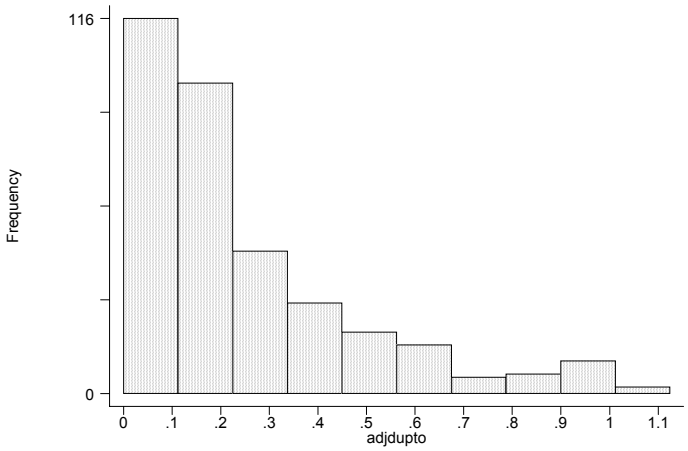


Table 11 **AQoL Population Norms, Disutility**

Age	AQoL (LD)			
	Male	n	Female	n
<25	0.02	5	0.05	5
25-34	0.10	24	0.14	38
35-44	0.06	21	0.14	71
45-54	0.20	20	0.21	64
55-64	0.34	18	0.27	27
65-74	0.20	15	0.20	18
Age >75	0.21	8	0.23	7

Table 12 **Population norms AQoL Disutility, PTO Disvalue**

Age	AQoL Total	PTO Total
<25	0.04	0.08
25-34	0.13	0.20
35-44	0.12	0.19
45-54	0.21	0.28
55-64	0.36	0.36
65-74	0.20	0.28
>75	0.22	0.31

Table 13 **PTO Population Norms (Dis Value)**

Age	AQoL (LD)			
	Male	n	Female	n
<25	0.07	5	0.09	5
25-34	0.18	24	0.21	38
35-44	0.12	21	0.21	71
45-54	0.27	20	0.29	64
55-64	0.41	18	0.33	27
65-74	0.28	15	0.29	18
Age >75	0.36	8	0.33	7

Table 14 **Standardised PTO Values**

Age	Male	Female	Total
<25	1.00	1.00	1.00
25-34	0.77	0.78	0.79
35-44	0.89	0.75	0.79
45-54	0.81	0.67	0.68
55-64	0.52	0.62	0.60
65-74	0.64	0.61	0.65
>75	0.58	0.55	0.59

TTO - PTO Transformation - Results

The transformation above might be used in one of two ways. First, it might be employed to transform the average TTO result into an average PTO score. Secondly, individual TTO scores may be transformed and the resulting of PTO scores averaged.

These two procedures are not equivalent. This is illustrated in Figure 5. This represents a TTO-PTO transformation. Assume two individuals have TTO scores of T_1 and T_2 and, therefore, an average TTO score of T_3. The PTO scores corresponding to T_1 and T_2 are P_1 and P_2 respectively which (because of the geometry) have an average value of P_{ave}. In contrast, the average TTO score of T_3 has a higher TTO score of P_3. In sum, the average PTO score is different from the score of the average TTO response. For all of the purposes of summary measures, it is the former which is required. Decisions are based upon average population scores and not upon the score of a constructed average person. For this reason we present below average PTO scores after transforming each individual's TTO into a corresponding PTO value.

Figure 5 Average of utilities versus utility of the average

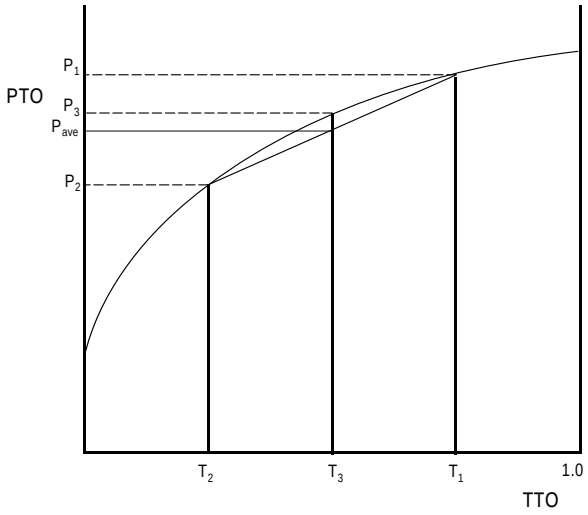
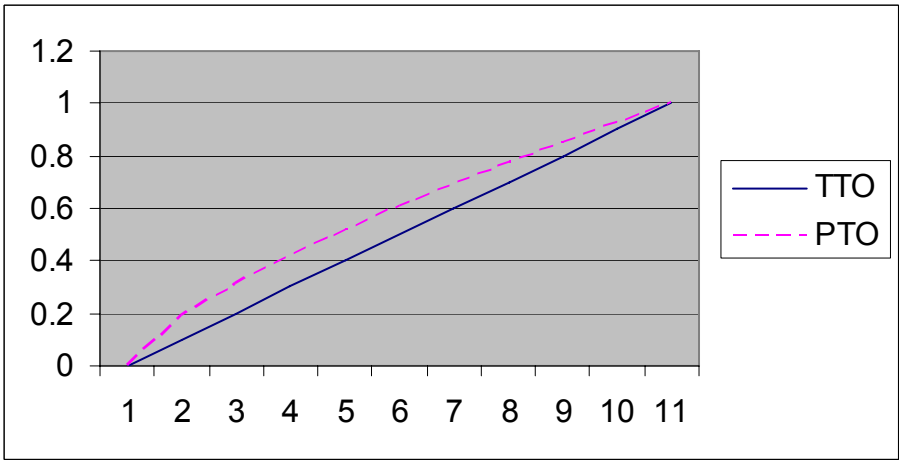


Figure 6 Transformation of TTO into PTO values



References

Fearon JD 1998, Deliberation as discussion. In Elster, J. (ed.) *Deliberative Democracy*, 44-68. Cambridge: Cambridge University Press.

Murray C and Lopez A 1996, *The Global Burden of Disease,* Geneva: WHO. Harvard University Press.

Nord E, Street A, Richardson J 1995, 'The significance of age and duration of affect in social evaluation of health care', *Health Care Analysis*, vol 4, pp 103-111.

Nord E, Richardson J et al 1995, 'Who cares about cost; does economic analysis impose or reflect social values?', Health Policy, vol 34, pp79-94.

Nord E, Richardson J et al 1995, 'Maximising health benefits versus egalitarianism: an Australian survey of health issues', *Social Science and Medicine*, vol 41, no 10, pp 1429-1437.

Nord E 1999, *Cost Value Analysis*, Cambridge University Press, Cambridge.

Richardson J & Nord E 1997, 'The importance of perspective in the measure of quality adjusted life years', *Medical Decision Making*, vol 17, pp 33-41.

Richardson J 1994, 'Cost utility analysis: what should be measured', *Social Science and Medicine*, 39:(1)7-21.

Sheill A, Seymour J, Hawe P and Cameron S 2000, 'Are preferences over health states complete?', *Health Economics*, 9, 47–55.

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

Appendix 1 AQoL Interviewer Instructions

PTO Instructions

INTRODUCTION

Explain the purpose of the PTO exercise in Session II:

In the first session the participant was asked to answer questions based on **his/her own imagined health** states with himself/herself as **RECEIVER of health services *(TTO questions)***. He/she was asked to make a choice between a shorter, healthier life versus a longer but less healthy life.

We are now going to ask the participant to answer questions based on imagined **health states for other people** with himself/herself as **PROVIDER of money for health services *(PTO questions)***. We would like them to choose between a number of people getting a treatment for a given illness/condition versus a number of people getting a different treatment for a different illness/condition.

A SIMPLE EXAMPLE

This is OPTIONAL. It can consume a lot of time, raise many questions than can interrupt the flow of the interview and lead to a confusing start, for both informant and interviewer. After the first interview, interviewers should be in a position to assess who would benefit from it, who doesn't need it and who would be confused further by it.

Present the PTO board to the participant using the **Simple Example PTO** sheet.

First, ask them to recall the example from the first interview (walking to a bar vs. staying in seat during intermission).

We are now going to ask them to imagine that they are the manager of the cinema deciding what type of bar or café to provide for customers. As manager, they have noticed that **customers ALREADY ARE both hungry and thirsty** by the intermission, and this spoils their enjoyment of the evening.

(a) **Option A** uses the money to build a café to provide seating for 50 people to enjoy a snack and a drink during the intermission while seated, ie **provides both food and drink**. These 50 people return to their seats satisfied (content) for the rest of the evening.

(b) **Option B** uses the money to build a bar which **provides a drink while standing**. People return to their seats for the rest of evening, but are not fully satisfied – they are no longer thirsty, but are still hungry.

That is:

(a) Option A "treats" hunger and thirst for 50 people which makes them content, by providing food and drinks

(b) Option B "treats" thirst but not hunger for a number of people which makes them partly content, by providing drinks only and leaving them hungry.

Only one of Option A or Option B can be provided – money is not an issue. The cinema holds 700 people.

Explain to the participant that we are going to ask them how many people in Option B would be needed to ensure that options A & B **give equivalent benefits to the patrons**. There are no right or wrong answers, we are interested in their views and opinions.

Explain the principle of the PTO board and the People Cards:

(a) Spread the People Cards on the table beside the PTO board. Keep them in groups according to the number of people of each card, but don't arrange the groups of cards in any particular order.

(b) Place a 50 People Card under Option A and a 50 People Card under Option B.

(c) If you could provide Option A which "treats" hunger and thirst for 50 people or Option B which "treats" thirst for 50 people, which would you choose?

(d) The participant should say Option A. Add Cards under Option B so that the total becomes 250 people, and ask them to choose again

(e) Follow the principles of flip-flopping by adding and taking away People Cards under Option B until they say Option A and Option B are equivalent. Always leave the 50 People Card under Option A.

PTO QUESTIONS

Go to the PTO sheets

Briefly recap for the participant that:

(a) they are to imagine they are a decision-maker in the Australian Health Department
(b) they have to decide which treatments that improve health are provided
(c) different treatments will improve health in different ways and can benefit different numbers of people
(d) they have to choose between treatments because the country cannot afford to provide all treatments for all conditions

Emphasize

--no right or wrong answers.
--their own personal values and the importance they would put on different poor health states.
--we are not trying to change their minds or educate them; the usefulness and validity of the study depends on how well their answers reflect their own personal values – do this as BEST AS THEY ARE ABLE.
(If you find that you are working too hard to get an answer, back off and keep reminding the informant that the researchers are really interested in how they think and what values they hold about health.)

Explain that we are now going to ask them to **choose between two treatments which improve health for people with two different conditions**.

Ideally participants should complete all type C, D and E questions, as per their PTO Score sheet.

Notes:

(a) **If Health State B is worse than death, participants should choose A as their response**. That is, if Health State B is worse than death they would not treat any people with Treatment B, they would rather let them die. The informant believes that condition B is so bad that death would be preferable.

Where participants indicate Health State B is worse than death, go to the Worse Than Death PTO question.

(b) **If Health State B is not worse than death, participants should choose any number from 100 upwards**. They do not have to stick to choosing numbers shown on the People Cards – they are just a guide. Use the Cards to get participants to "add up" numbers for Treatment B.

If they choose 100, this implies Health State B is equally as good as Excellent Health. If they choose a number greater than 100, this implies Health State B is poorer than Excellent Health. The higher the number, the worse the informant believes the condition is. The money spent on treatment B would have to save many (more than 100) people, even if they end up in a "condition", to justify not spending it to restore 100 people to excellent health, ie "condition-free".

Of course, as we have found out, this can raise many issues, such as “I think it would be cruel to save more people only to have them live in such a dreadful state.” “I can’t bear to think of being responsible for more people like that living in the world.” etc.

Remember it’s their value we are interested in, whatever they think.

If the participant wants to go back and change answers later, allow them to. If they are repeatedly revising answers to earlier questions suggest they complete all questions and then go back.

C, D AND E TYPE PTO SCRIPT

We want you to think about different treatments that can improve health for people with two different conditions. Without treatment, people suffering from **either condition** will die within 3 months. We want you to think about two treatments:

A. 100 people in your country have a rapidly fatal condition which can be treated with **Treatment A**. The identity of these 100 people is unknown. **Without Treatment A** these 100 people will die suddenly in the next 3 months (Point to the top box – Immediate Death). **With Treatment A** these 100 people will live in excellent health (Point to Health State A – Excellent Health) for a normal life time.

B. A number of different people in your country have a different rapidly fatal condition which can be treated with **Treatment B**. The identity of these people is unknown. **Without Treatment B** these people will die in the next 3 months (Point to the top box – Immediate Death). **With Treatment B** these people will live in poor health (Point to Health State B) for a normal life time.

The people who would benefit from Treatment A or Treatment B represent a broad cross-section of the whole population, and there are no differences in terms of age, sex, race, socio-economic status etc. *(may not need to repeat this section after first couple of questions).*

From a quality of life point of view, how many people getting Treatment B would **they consider equivalent** to 100 people getting Treatment A? Flip-flop starting at 100 with the People Cards:

(a) Spread the People Cards on the table beside the PTO board. Keep them in groups according to the number of people of each card, but don’t arrange the groups of cards in any particular order.

(b) Place a 100 People Card under Treatment A and a 100 People Card under Treatment B. If you could provide Treatment B for 100 people, which would Treatment would you choose?

(c) Follow the principles of flip-flopping by adding and taking away People Cards under Treatment B until they say Treatment A and Treatment B are equivalent. Always leave the 100 People Card under Treatment A.

The higher the number the worse they consider the health state to be.

If Health State B is worse than death, participants should choose A as their response, and then go to the Worse Than Death PTO question (C, D type only).

If Health State B is not worse than death, participants should <u>choose any number from 100 upwards</u>, and then skip the Worse Than Death PTO question.

Record their answer.

Go through the remainder of the questions till all are answered.

TAKE A SHORT BREAK

INTERVIEWER PROMPTS FOR PTO INTERVIEWS

Today we are going to be talking about Quality of Life for others instead of ourselves. Here we have (*point to black square Without Treatment*) a number of people who are suffering a disease that will see them die within the next three months.

Point to Treatment A

We have 100 of these people who will receive Treatment A and they will be returned to Excellent Health and live in that condition for the remainder of their life.

Point to Treatment B

We have another 100 of these people who will receive Treatment B and they will be returned from Immediate death but will live in Poor health (point to health state) for the rest of their lives.

From a Quality of Life point of view do you think the 100 people receiving Treatment B and living in this state for the rest of their lives, is equal to the people receiving Treatment A who are returning to excellent health for the rest of their lives. Does the scale (point to scale) sit balanced for you?

If the respondent says no, offer them the people cards and ask them to add to the 100 people in Treatment B until they think the scales are even.
If the number keeps increasing (say to 2000 and more), ask them if death would be a release. If yes, then inform then that that is worse than death.

Worse that Death question following PTOT C1-6 questions

If the respondent chooses the Worse than death answer –

Place 100 people above Treatment A and explain that these people were living in this chronic condition (that they indicated was worse than death) for the rest of their lives but will receive Treatment A which will return them to excellent health.

Place 100 people above Treatment B and explain that these people were in suffering a terminal illness that would see them die within the next three months, but will receive Treatment B which will return them to excellent health.

From a Quality of Life point of view do you think the 100 people living in this chronic condition (point to Health State A) who will receive Treatment A and go on to live in excellent health for the rest of their lives, is equal to these 100 people facing immediate death (point to Health State B) who will receive Treatment B and go on to live in excellent health for the rest of their lives.
Does the scale (point to scale) sit balanced for you?

If the respondent says no, offer them the people cards and ask them to add onto the 100 people in Treatment B until they think the scales are even.

C & D TYPE WORSE THAN DEATH PTO SCRIPT

We now want you to think some more about this Health State. This time, we want you to think about two different treatments which prevent poor health for people with two different conditions:

A. 100 people in your country will get a chronic condition which can be prevented with **Treatment A**. The identity of these people is unknown. **Without Treatment A** these 100 people will live in very poor health (Point to Health State A) for a normal life time. Health State A is the Health State you told me you consider to be worse than death. **With Treatment A** these people will live in excellent health (Point to the top box) for a normal life time.

B. A number of people in your country will get a rapidly fatal condition which can be prevented with **Treatment B**. The identity of these people is unknown. **Without Treatment B** these people will die in the next 3 months (Point Health State B – Immediate Death). **With Treatment B** these people will live in excellent health (Point to the top box) for a normal life time.

The people who would benefit from Treatment A or Treatment B represent a broad cross-section of the whole population, and there are no differences in terms of age, sex, race, socio-economic status etc. *(may not need to repeat this section after first couple of questions).*

From a quality of life perspective, how many people getting Treatment B would **they consider equivalent** to 100 people getting Treatment A? (flip-flop starting at 100 using the People Cards).

(a) If Health State A is worse than death, participants should choose any number from 100 upwards.

(b) If Health State A is equally as bad as death, they should choose 100.

Record their answer.

Go on to the next question.

Appendix 2 Visual Aid for PTO Questions