

**OUTCOME VALUATION IN THE ECONOMIC
EVALUATION OF HEALTHCARE**

by

Richard P.A. NORMAN

Submitted to the University of Technology, Sydney

for the degree of Doctor of Philosophy

Submitted October, 2012

CERTIFICATE OF AUTHORSHIP / ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Acknowledgements

My first thanks are to my panel of supervisors, Professor Jane Hall, Associate Professor Rosalie Viney and Professor Debbie Street. Jane has offered me both excellent support and a critical eye, which has helped immeasurably in the development of the thesis. As Director of the Centre for Health Economics Research and Evaluation (CHERE), she has established an environment in which junior researchers are encouraged to pursue their own research agenda, something which has enriched my time so far in the department. Rosalie was responsible for my initial interest in the field, and I have enjoyed working with her, particularly on our successful NHMRC Project Grant, and look forward to continuing the collaboration into the future. Her knowledge of the field is invaluable, and she has consistently helped me reframe my arguments into something more coherent and convincing. Debbie has shown incredible patience as I asked the same subset of muddled questions regarding the mathematics behind designing experiments, consistently responding with humour and advice, both of them good. As a team, the panel has worked seamlessly, and I think that the thesis would be much diminished without the participation of any one of them.

I want to acknowledge the financial support I received through my candidature. I received a PhD scholarship from the Centre for the Study of Choice (CenSoC). Additionally, I received a top-up scholarship from CHERE. The provision of both was essential to allow me to pursue this study and I am eternally grateful that both organisations supported me in this way. I am also grateful to the University Graduate School for providing support to attend the International Health Economics Association Congresses in 2009 and 2011 to present elements of this work.

I want to thank Survey Engine, particularly Ben White, for support in the running of the two experiments reported in the empirical chapters. The academic rates they provided for both experiments were much appreciated, as was their overseeing of the data collection process. Similarly, I would like to acknowledge the time and effort of the anonymous survey respondents, recruited through PureProfile, without which the thesis would not have been possible.

I want to thank the staff and visitors at CHERE past and present for their insight and feedback in the various seminars in which I have presented this work. I would also like to acknowledge the input of Dr. Leonie Burgess, who kindly provided the design used in the SF-6D experiment reported in Chapter 5. I would like to thank Liz Chinchon for performing the literature searches in the thesis and for proof-reading the final draft. I want to thank Professor Denzil Fiebig for kindly allowing me to attend courses at the University of New South Wales. The econometrics and discrete choice data modelling courses proved essential in completion of the thesis.

Finally, I want to thank Charmaine, Luke and Isla. Since the time I started discussing the possibility of undertaking this PhD within CHERE, I have managed through incredible good fortune to gain a wife, a son, and a daughter who have supported me with love, kindness and endless demands for re-runs of Toy Story (the last one was mostly Luke). Over the three years, Charmaine has been endlessly willing to ask how it was all going, and more importantly, to graciously endure the answer.

Table of Contents

OUTCOME VALUATION IN THE ECONOMIC EVALUATION OF HEALTHCARE.....	i
Acknowledgements.....	iii
Table of Contents.....	iv
Figures.....	vii
Tables.....	viii
Appendices.....	ix
Abstract.....	x
Chapter 1: The measurement of outcomes in economic evaluation of health interventions.....	1
Chapter Summary.....	1
Economic evaluation in healthcare.....	1
Arrow’s characterisation of the healthcare market.....	3
A Welfarist underpinning of economic evaluation in healthcare.....	4
Pareto and Kaldor-Hicks criteria.....	6
The Utility Principle, Individual Sovereignty and Consequentialism.....	8
Cost-Benefit Analysis.....	9
Extra-Welfarism.....	10
Respecifying the desideratum.....	11
The roots of Extra-Welfarism.....	12
The Quality-Adjusted Life Year – A history and critique.....	15
Evaluating the QALY.....	18
Using the QALY – A problematic example.....	19
Individual preference constraints in the construction of the QALY model.....	20
Some additional criticisms of extra-welfarism.....	24
Beyond Welfarism and Extra-Welfarism.....	25
Communitarianism.....	25
Empirical Ethics.....	26
Initial conclusions.....	29
Thesis structure.....	30
Chapter 2: Measuring health-related quality of life – standard and novel approaches	32
Chapter summary.....	32
Introduction.....	32
Section A: Methods for valuing health states.....	33
Standard Gamble.....	33
Time Trade-Off.....	36
Visual Analogue Scales.....	40
Section A Conclusion.....	41
Section B: Multi-Attribute Utility Instruments.....	41
Categorising approaches to describing and valuing quality of life profiles.....	42
Whose values?.....	43
A framework for building and evaluating MAUIs.....	44
Structural independence.....	45
EuroQoL - 5 Dimensions (EQ-5D).....	46
Short Form – 6 Dimensions (SF-6D).....	51
Health Utilities Index (HUI).....	58
Assessment of Quality of Life (AQoL).....	59
Section B conclusion.....	61

Section C: Imputing values for other health states	61
Parametric approaches	62
Additivity (e.g. EQ-5D and SF-6D).....	64
Multiplicativity (e.g. Health Utilities Index (HUI)).....	66
Chapter conclusion.....	68
Chapter 3: Discrete Choice Experiments: Principles and Application for Health Gain	70
Chapter summary	70
Introduction.....	71
Stated and revealed preference data.....	71
The role of Random Utility Theory	73
A suitable numeraire	74
Choice experiments and health gain	76
Lexicographic preferences surrounding death.....	78
Modelling respondent heterogeneity.....	79
Observable characteristics and heterogeneous responses	79
Modelling heterogeneity on unobservable characteristics.....	80
Conditional logit modelling (heterogeneity exploration model 1)	81
Base case analysis - Random-effects (RE) modelling	82
Heterogeneity exploration model 2 - Scale Multinomial Logit modelling.....	83
Heterogeneity exploration models 3 and 5 - Mixed logit analysis	84
Heterogeneity exploration models 4 and 6 - Generalised Multinomial Logit modelling	87
Two brief computational issues	91
Model evaluation	91
Deriving welfare measures from discrete choice experiments	92
The Hicksian Compensating Variation.....	93
Marginal rates of substitution	95
Using a ratio of marginal utilities	96
Chapter summary	97
Chapter 4: Some Principles for Designing Discrete Choice Experiments.....	99
Chapter summary	99
DCE design principles	99
Introduction to design theory	99
Contrasts and fractional factorial designs	102
The likelihood function and maximum likelihood estimators	111
Deriving the information (Λ) matrix.....	113
<i>B</i> and <i>C</i> -matrices.....	116
D-efficiency	118
Alternatives to D-efficiency.....	119
Design strategies	121
SAS Algorithms	128
Some other areas of interest.....	129
Chapter summary	131
Chapter 5: Using a Discrete Choice Experiment to Value Health Profiles in the SF-6D	132
Chapter summary	132
Introduction – Using ordinal data to value health states.....	132
Applications of DCEs to value health profiles	133
The SF-6D.....	135

The vitality dimension	137
Implausibility of health states	138
Design and presentation of experiment.....	138
Data and sample recruitment	140
Analysis.....	141
Non-linearity in the utility function (models D and D1-D6)	143
Limitations in specifying random parameters.....	145
Rescaling scores for economic evaluation	145
Additional sub-group analysis	146
Results.....	148
Marginal frequencies	150
Base case utility weights for the SF-6D.....	156
Sub-group analysis.....	158
Heterogeneity modelling.....	162
Utility Function A	162
Utility function D.....	168
Overall model comparisons	174
Deriving utility weights under models A1-A6.....	177
Chapter discussion	179
Chapter 6: Equity Weights for Use in Economic Evaluation	186
Chapter summary	186
Introduction to equity in economic evaluations of health care interventions	186
Equity and altruism	187
Social Welfare Functions and equity	188
Criticisms of SWF linearity	190
Symmetry of the SWF	191
Identifying relevant literature	193
Existing attempts to estimate a SWF using stated preference data.....	194
Identifying dimensions for the DCE.....	200
Including gender	201
Age weighting.....	202
Life expectancy, current age or both?.....	203
Selecting dimensions and levels for the DCE.....	204
Designing the choice experiment.....	206
Sample recruitment	208
Analysis.....	208
Relaxation of the utility function (models B, B1-B6).....	209
Self-interest and empathy – sub-group analysis	210
Generating equity weights from regression results.....	211
Results.....	214
Marginal frequencies	215
Random-Effect probit results.....	217
Heterogeneity based on observed respondent characteristics	220
Modelling heterogeneity	223
Utility function A	223
Utility function B	227
Model comparison	231
Generating equity weights	234
Conclusions and implications	238
Chapter 7: Conclusions and Implications	243

The state of economic evaluation of healthcare.....	243
The description and valuation of health.....	244
Discrete choice experiments	245
DCE 1 – Valuing the SF-6D health states	246
DCE 2 – Equity weights for economic evaluation.....	247
Summary of importance.....	248
Some future directions	249
Appendices.....	251

Figures

Figure 1: The Kaldor-Hicks Criterion vs. Pareto Criterion	7
Figure 2: Alternative Health Profiles over Time	17
Figure 3: An Unconstrained Social Welfare Function.....	27
Figure 4: A Constrained Social Welfare Function.....	28
Figure 5: The Standard Gamble	34
Figure 6: The Time Trade-Off for states considered better than death.....	37
Figure 7: The Time Trade-Off for states considered worse than death	37
Figure 8: Health Visual Analogue Scale.....	40
Figure 9: Comparing UK results with other leading studies.....	50
Figure 10: Self-Assessed Health Using the EQ-5D and the SF-6D.....	56
Figure 11: Nesting Regression Models Within the G-MNL.....	90
Figure 12: The Compensating Variation (CV)	94
Figure 13: Ratio of Marginal Utilities	96
Figure 14: Flowchart of algorithm for constructing efficient choice designs.....	129
Figure 15: An Example Choice Set	139
Figure 16: SF-6D Dimension / Level Marginal Frequencies.....	150
Figure 17: RE probit and logit coefficients (Model A).....	155
Figure 18: RE probit and logit coefficients (Model D).....	155
Figure 19: Distribution of SF-6D health states (corrected utility function 1, random-effects probit).....	158
Figure 20: Sub-Group Analysis Results (Gender of respondent)	159
Figure 21: Sub-Group Analysis Results (Age of Respondent).....	160
Figure 22: Sub-Group Analysis Results (Chronic Conditions)	161
Figure 23: Comparison of Akaike Information Criteria (AIC).....	175
Figure 24: Comparison of Bayesian Information Criteria (BIC) (n=observations)...	175
Figure 25: Comparison of Bayesian Information Criteria (BIC) (n=individuals).....	176
Figure 26: Comparison of Health State Valuation under Different Algorithms.....	181
Figure 27: Comparison of utility weights associated with general population sample using pre-existing SF-6D and EQ-5D algorithms.....	183
Figure 28: Comparison of utility weights associated with general population sample using Australian DCE-derived algorithms.....	184
Figure 29: Symmetrical Utilitarian and non-Utilitarian Social Welfare Functions...	189
Figure 30: Relaxing the symmetrical assumption in non-linear SWF's	192
Figure 31: A set of symmetrical SWFs with constant elasticity of substitution assuming anonymity (i.e. $\alpha = \beta$)	196
Figure 32: A set of SWFs with constant elasticity of substitution allowing differing interpersonal weights	196
Figure 33: An Example Choice Set	206

Figure 34: Marginal Frequencies	216
Figure 35: Comparison of Coefficients under Utility Function A.....	219
Figure 36: Comparison of Coefficients under Utility Function B	219
Figure 37: RE Probit sub-group analysis (gender)	221
Figure 38: RE Probit sub-group analysis (smoking).....	222
Figure 39: RE Probit sub-group analysis (carer status)	223
Figure 40: AIC figures for the 12 Models	232
Figure 41: BIC figures for the 12 Models (n=individuals).....	233
Figure 42: BIC figures for the 12 Models (n=observations)	234
Figure 43: Distribution of Equity Weights	238

Tables

Table 1: The EQ-5D.....	46
Table 2: Self-Assessed Health (EQ-5D) (n=2,494)	48
Table 3: Correlation coefficients between self-assessed EQ-5D dimensions.....	48
Table 4: Existing EQ-5D Algorithms	49
Table 5: The SF-6D	52
Table 6: SF-6D Self-Assessed Health (n=2,494).....	54
Table 7: Correlation Coefficients between self-assessed SF-6D dimensions.....	54
Table 8: HUI3 Multi-Attribute Utility Function.....	68
Table 9: Contrasts for main effects in a 2^3 experiment.....	103
Table 10: Contrasts for main effects and interactions in a 2^3 experiment	104
Table 11: (Non-orthogonal) contrasts for main effects in a 3^3 experiment	105
Table 12: A, B, and AB contrasts for main effects and interactions in a 3^3 experiment	106
Table 13: A 2^{5-1} fractional factorial design.....	107
Table 14: Non-overlapping regular designs.....	109
Table 15: Example L^{MA} Design	123
Table 16: Example Main-Effects Only Choice Experiment	124
Table 17: Selecting generators to estimate main effects and interactions	127
Table 18: The SF-6D	135
Table 19: The Vitality Dimension	138
Table 20: Models Run in Chapter 5	144
Table 21: Representativeness of SF-6D DCE Sample.....	149
Table 22: Sample SF-6D Health (n=1,017)	150
Table 23: Results From Models A-D.....	152
Table 24: Base case QALY algorithm	157
Table 25: Information Criteria (Gender Sub-Group Analysis).....	159
Table 26: Information Criteria (Age Sub-Group Analysis).....	160
Table 27: Information Criteria (Chronic Conditions Sub-Group Analysis)	162
Table 28: Heterogeneity Modelling Specification Results (Utility Model A).....	163
Table 29: Variance-Covariance Matrices for Model A5	167
Table 30: Variance-Covariance Matrices for Model A6	168
Table 31: Heterogeneity Modelling Specification Results (Utility Model D).....	169
Table 32: Variance-Covariance Matrix for Model B5.....	173
Table 33: Variance-Covariance Matrix for Model B6.....	174
Table 34: Model Comparison	174
Table 35: DCE-derived QALY Weights for the SF-6D (Main Effects Only).....	178
Table 36: Correlation Coefficients for the 18,000 Health State Valuations	179

Table 37: Spearman Rank Coefficients for the 18,000 Health State Valuations.....	179
Table 38: Agreement between instruments under existing and novel methods.....	185
Table 39: Potentially relevant personal characteristics identified by Olsen <i>et al.</i>	200
Table 40: Dimensions and levels for the choice experiment	205
Table 41: A starting design of 2 ⁵ in 16 rows (strength 4)	206
Table 42: Models Run in Chapter 6.....	210
Table 43: Representativeness of DCE Sample	215
Table 44: RE Probit Results.....	217
Table 45: RE Logit Results.....	218
Table 46: Heterogeneity Modelling Results (Utility Function 1).....	225
Table 47: Heterogeneity Modelling Results (Non-Linear Utility Function)	228
Table 48: Model Comparison	231
Table 49: Equity Weights	235
Table 50: Equity-Efficiency trade-off search strategy.....	242

Appendices

Appendix 1: HUI Mark 3	252
Appendix 2: The Assessment of Quality of Life instrument	255
Appendix 3: Final SF-6D DCE Design	258
Appendix 4: SF-6D DCE Screen Shots	262
Appendix 5: RE Probit and RE Logit Results under a Non-Linear Utility Function	272
Appendix 6: SF-6D DCE Subgroup Analysis (Gender).....	274
Appendix 7: SF-6D DCE Subgroup Analysis (Age)	276
Appendix 8: SF-6D DCE Subgroup Analysis (Chronic Conditions)	278
Appendix 9: Equity Weights Experiment	280
Appendix 10: Equity Weights for Economic Evaluation DCE	283
Appendix 11: Equity Weights gender subgroup analysis	290
Appendix 12: Equity Weights smoker subgroup analysis	293
Appendix 13: Equity Weights carer status subgroup analysis.....	295
Appendix 14: Variance Covariance Matrices (Utility Function A).....	297
Appendix 15: Variance Covariance Matrices (Utility Function B).....	299

Abstract

Economic evaluation of healthcare interventions (such as pharmaceuticals, medical devices and technologies) considers both the effect of the intervention on patients, and the costs borne by the government and often the individual themselves. This simultaneous consideration of costs and benefits is now standard practice in reimbursement decisions, both in Australia and elsewhere. This thesis focuses on the assessment of benefits, specifically how we place a value on the health changes patients experience as a result of a health care intervention.

There is a well-established framework for how outcomes are valued in health care, but this framework is built on a number of contentious assumptions. For example, health is assumed to be the sole outcome of a healthcare system, and society is assumed to be inequality-neutral. This thesis identifies and explains these assumptions and then focuses on testing two of them in the empirical chapters. The overall aim of the thesis is to explore the extent to which the current framework reflects population preferences, and whether the framework can be adapted to be more reflective of population preferences. The empirical chapters in this thesis consider these issues, using a discrete choice experiment (DCE). For reasons presented in Chapters 3 and 4, this technique offers very attractive properties for answering these types of questions.

The standard approach to valuing health outcomes uses the quality-adjusted life year, in which the value of a health profile is the product of quality of life and length of life. For this to be operationalised, we need to be able to describe health states in a way which captures all relevant dimensions of quality of life that are important to people, and then we need to assign values to health states. This thesis argues that the current methods for assigning values to health states are very onerous for survey respondents, and prone to significant bias. Standard valuation techniques require the respondent to identify preferences around quality of life through the acceptance of a risk of death, or the reduction of life expectancy to alleviate poor quality of life. However, these fail to control for issues such as risk-aversion or time preference. The first empirical analysis uses a DCE to value health states for the SF-6D, a health state valuation instrument that is based on the very widely used quality of life instrument the SF-36. The use of a DCE aims to remove (or control for) these biases. This chapter represents a

methodological advance through the use of a DCE, and produces the first Australian algorithm for the SF-6D.

The second empirical analysis considers the assumption that the value of health improvement is independent of who receives it. Therefore, it is conventional for an extra year in full health to be regarded as being of the same value to society independent of who receives it. The chapter results suggest that the average respondent prefers giving additional health to people with low life expectancies, carers, and non-smokers even if it reduces total health for society as a whole. The chapter concludes by identifying how these preferences might be integrated into economic evaluation.

This thesis explores two areas in which the conventional approach to outcome valuation in economic evaluation are subject to concern. It demonstrates how these concerns might be overcome by augmenting the existing framework with relatively easily-collected stated preference data, and offers a template for other analyses exploring other parts of how health outcomes should be valued.

Chapter 1: The measurement of outcomes in economic evaluation of health interventions

Chapter Summary

In this chapter, the concept of economic evaluation in healthcare is introduced. Firstly, a justification for societal intervention in resource allocation in health is considered, and it is concluded that there are valid reasons for moving away from a *laissez-faire* approach. The ways in which economic evaluation differs in health from other areas where economics plays a role are then discussed. In particular, the move away from a utility-centric model is evaluated. The dominant extra-welfarist approach is introduced, in which health is decoupled from utility, and is considered to be the maximand. The usual metric used in this extra-welfarist framework, the quality-adjusted life year (QALY) is then discussed, both in terms of how it developed, and also areas in which it may diverge from many people's concepts of the outcomes of a health intervention. Then, a number of alternatives to welfarism and extra-welfarism are introduced. The chapter concludes with an outline of the structure of the thesis.

Economic evaluation in healthcare

Decisions about health and healthcare are difficult. Individuals are constantly making decisions that have potential implications for their future experience of good or poor health. While personal activity is likely to be a significant determinant of health for most people, society plays a significant role in the health of its members also. As a society, we have an infinite number of ways in which we can spend money on health, but only a finite budget. Therefore, choices have to be made. Choosing between, for example, expanding a neonatal ward in a hospital and a public health intervention targeting obesity is likely to be emotive and to involve a variety of considerations. Do we pick the option which saves the most lives? How do we choose between an option which save lives, and one which improves quality of life? Do we value health of neonatal infants differently because of who they are? Is obesity a reason for prioritising the health of an individual differently? Does the cost of each intervention matter?

In a resource constrained environment, these kinds of decisions need to be made. They may not be as stark as choosing between directly competing options, but

ultimately there must be a way of valuing a particular healthcare intervention so those interventions which are in some sense ‘value for money’ can be identified. It is advantageous to have the assumptions upon which this decision making process exists explicit and acceptable to society. It is this idea, of forming a justifiable framework where decisions can be made rationally, which underpins a formal process to evaluate possible use of health resources across most developed countries (for instance see examples in England and Wales (National Institute for Health and Clinical Excellence, 2008; National Institute for Health and Clinical Excellence, 2007), in Canada (Canadian Agency for Drugs and Technologies in Health, 2006), and in Australia (Department of Health and Ageing, 2007; Department of Health and Ageing, 2005)). While the approaches taken in different countries are tailored to reflect the unique circumstances in which healthcare decisions are made, these types of approaches will typically include considerations of safety, effectiveness, and cost-effectiveness.

A fundamental question which requires addressing is this: Why do societies have to intervene regarding healthcare resource allocation? As a society, there are many areas where we do not impose government control over allocation. It is standard in economics to assume that individuals have perfect information regarding the markets they enter, and, if there is no societal intervention, will make decisions that maximise their welfare. If everyone’s welfare is maximised, the welfare of society is maximised. Thus, if we as a society provide a resource to someone who would have received it under a free-market allocation of resources, their benefit from receiving it must be less than the cost of provision (and the opposite argument can be made for limiting access to a resource). If we are to place constraints on how the market for health resources works, we need to know what is different about health to make this government intervention appropriate. Might we argue that, in the health sector, individuals make decisions which do not maximise their own health? Maybe we could justify societal intervention by arguing against the proposition that individuals maximising welfare maximise societal welfare? These are difficult arguments to support and, to investigate them fully, it is necessary to outline what is meant by welfare analysis, and then to consider some of the reasons why health might be special.

In welfare analysis, social welfare is a function of the welfare of each of the individuals within the society (this will be discussed later in this chapter). In a perfectly competitive market, each individual chooses the option that maximises his or her own welfare, and this means that the welfare of society is maximised (under the assumption that social welfare and individual welfare are defined in the same dimension). However, the market for health may be subject to failure, in the sense that a *laissez-faire* approach does not lead to a maximisation of social welfare (which can be termed as a sub-optimal allocation of resources). If this is true, the role of government might be to attempt to counteract this failure, by putting in place policies which maximise societal welfare. The major role of economic evaluation in healthcare is to mimic or replace perfect market allocation of resources. The necessity of it increases with the degree to which the health sector moves away from the assumptions required for perfectly competitive markets, under which welfare is maximised. These assumptions are now discussed in the context of Arrow's exposition of the healthcare system (1963).

Arrow's characterisation of the healthcare market

Regarding whether the health sector meets the criteria for these perfectly competitive markets, Arrow argued persuasively that it does not, and discussed the characteristics of the healthcare system that move it away from that which would maximise social utility without intervention (Arrow, 1963). The first identified source of market failure was that the nature of demand for health care was irregular and unpredictable. This means that consumers of healthcare have little experience of demanding healthcare, and cannot plan for future health expense accurately. This leads to issues surrounding insurance against future health expenditure and consequences of this, including moral hazard, in which an individual behaves differently (i.e. less cautiously) if protected financially from the consequences of a risk. Arrow's second source of market failure was that physicians have a role both as provider and inducer of demand for health care. This results from the asymmetric information between the physician and the consumer of healthcare. As the physician has greater knowledge of the area, and also because they act as gatekeeper to specialist care in some healthcare systems, the physician plays a key role in determining what healthcare their patient receives, and indeed may be the one to supply it. Either way, they are likely to have a financial incentive to act in a certain way. If we assume that (financial) self-interest on the part

of the physician plays any role at all, this can lead to an over-supply of services in which the marginal cost to society of additional provision of services exceeds the marginal benefit. Arrow's third source of market failure was the significant uncertainties associated with the expected outcomes from a medical service. The beliefs of the patient with regard to the expected welfare benefit from a service are based on very limited evidence, while the evidence for the effectiveness of a service across the entire population often shows considerable differences in individual responses to the same service. He then discussed the relatively high barriers to entry including professional licensing and the cost and time commitment required to become professionally qualified. Thus, agents in the market, and those who might enter it, are unable to promptly respond to changes in demand. Finally, he suggested that pricing practices differ in terms of price discrimination by income and a tendency towards fee for service. Arrow concluded that,

“(T)he failure of the market to insure against uncertainties has created many social institutions in which the usual assumptions of the market are to some extent contradicted. The medical profession is only one example, though in many respects an extreme one.”(p.967)

Therefore, in a health setting, the free-market is prone to lead to a distribution of outcome or resources which is in some sense sub-optimal. The definition of what constitutes optimal is difficult. However we decide to define the term, the government may choose to intervene to ensure a better allocation. Since Arrow described the unusual characteristics of the health sector, the need for proxying markets for health interventions or technologies has remained potent. The methods for correcting for market failure have been debated at length, and contentious issues abound in the literature.

A Welfarist underpinning of economic evaluation in healthcare

To this point, the emphasis has been on explaining why economic evaluation is preferable to a *laissez-faire* approach to the allocation of healthcare resources. In this section, the standard approaches used for outcome measurement in economic evaluation are outlined, specifically welfarism (which is widely used across economics sub-disciplines) and the health-tailored extra-welfarism which departs

from some of the assumptions of standard economic analysis¹. This will be subsequently contrasted with alternative non-welfarist approaches such as Communitarianism and Empirical Ethics.

Welfare economics is the evaluation of competing states of the world, and specifies a utility framework to enable ranking of these competing states from best to worse (Brouwer, et al., 2008). Welfare economics can be divided into two significant time periods, termed classical and neo-classical. It is the latter of these that most informs modern economic evaluation techniques. Thus, I will introduce classical welfare economics, and then discuss in detail the divergence between it and neo-classical welfare economics.

Classical welfare economics is characterised by Pigou, Edgeworth and Marshall, which draws heavily from the utilitarianism of John Stuart Mill. As Brouwer (2008) notes, classical welfare economics is based on welfarism, the cardinal measurement of utilities, and on the following additional characteristics,

“(i) The utility principle (i.e. individuals rationally maximise their welfare by ordering options and choosing the preferred option).

(ii) Individual sovereignty (i.e. individuals are themselves the best – some might say ‘the only’ – judge of what contributes most to their utility and how much that contribution is).

(iii) Consequentialism (i.e. utility is derived only from the outcomes of behaviour and processes rather than the processes themselves or intentions that led to the outcomes).” (p.327)

Neo-classical welfare economics departs from these assumptions by rejecting cardinality and therefore interpersonal comparability. Removing the former implies removing the latter because, if we cannot quantify a change in utility in two people under the same metric, we cannot then compare the sizes of utility change resulting from a change in resource allocation. Economists described as following the neo-classical welfare economics tradition include Pareto, Hicks and Kaldor.

¹ There is some inconsistency regarding the terms ‘Extra-welfarism’ and ‘Non-welfarism’. Extra-welfarism as a term will be used to refer to that of Culyer and Williams, with the emphasis on the QALY as the preferred outcome measure. Frameworks such as communitarianism and empirical ethics will be termed ‘Non-welfarist’.

Pareto and Kaldor-Hicks criteria

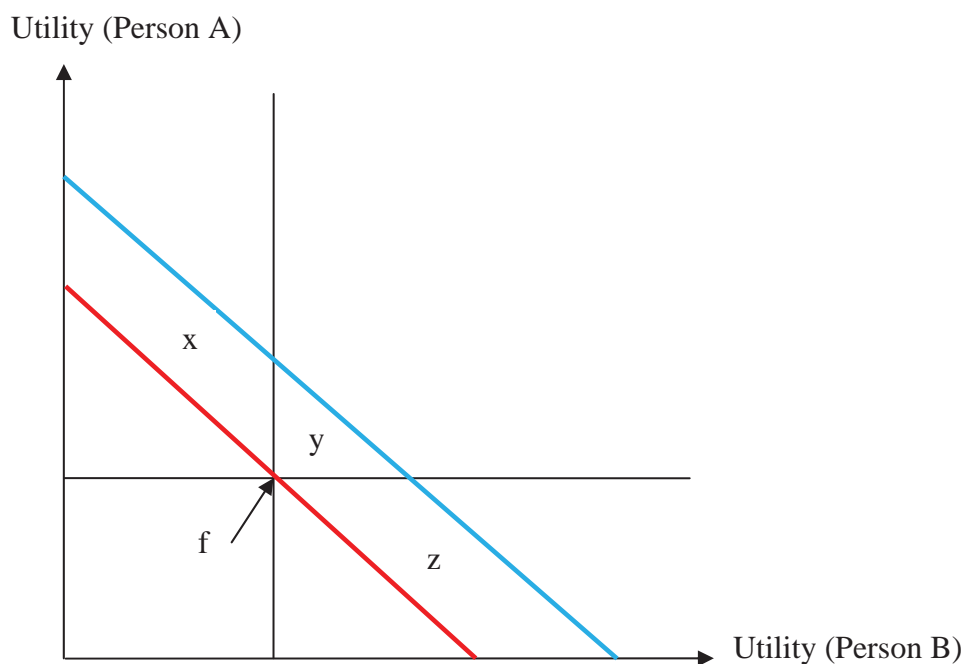
The removal of interpersonal comparability which is implied by moving from classical welfare economics to neo-classical welfare economics has important consequences. If we cannot compare outcomes accruing to different people, the obvious conclusion regarding the relative merits of different courses of action is that one is better than another only if it leads to a welfare improvement in at least one person, and a loss in welfare to no-one. This is because we have no way of balancing losses occurring to some groups and gains accruing to others. The ‘no-loser’ decision rule is called the Pareto criterion. Hurley (2000) echoes a range of authors by noting that strict enforcement of this comes at a high price:

“Because nearly all policy changes make someone worse off, strict application of the Pareto criterion leads to policy paralysis” (p.61)

In reality, such a strict implementation of the Pareto criterion is unlikely. Instead, a less restrictive criterion is employed in which a resource allocation is efficient if and only if the gainers from a move from the current allocation of resources can adequately compensate the losers and still gain overall (this is termed a potential Pareto improvement). This is known as the Kaldor-Hicks criterion, and can be explained diagrammatically, as shown in Figure 1. A crude Social Welfare Function is first defined, identifying points of equal total utility between two people A and B², who comprise all members of a society.

² A Social Welfare Function can be specified using non-utility measures. The Kaldor-Hicks criteria can apply if these non-utility measures are used instead; however as noted by Coast later in this chapter, it is problematic in situations where compensation is not possible.

Figure 1: The Kaldor-Hicks Criterion vs. Pareto Criterion



Imagine that this society is at point f on the red Social Welfare Function (SWF). The SWF is assumed to reflect perfect substitutability between utility for Person A and Person B: an unrealistic assumption but unimportant for identifying the notions of Pareto and potential Pareto improvement. Relative to the red SWF, the blue SWF consists of some points with a higher sum utility between A and B; indeed, as the two SWFs are parallel, the blue SWF reflects a fixed increase in aggregate welfare relative to the red SWF. Between the current red SWF and the better blue SWF, only those points in area y are Pareto improvements (as no-one is worse off, and at least one of the two people are better off). However, points in areas x and z are similarly improvements in terms of total utility. Thus, points in these areas meet the Kaldor-Hicks criterion (in which the gainer could compensate the loser), but do not meet the stricter Pareto criterion. Importantly, the use of the Kaldor-Hicks criterion is analogous to the maximisation of the total of the primary outcome (in this instance, utility). This measurement of value of gains and losses generally requires a monetary metric and is couched in terms of willingness to pay (or to accept). For a good defence of the Kaldor-Hicks criterion, see Harberger (1971).

There are three criticisms of the Kaldor-Hicks criterion that should be raised here. Firstly, it has been shown that it is possible for the Kaldor-Hicks criterion to be met

by a move from some allocation X to some other allocation Y, and also from Y to X (Scitovsky, 1941). Secondly, the value of a change in resources may depend on whether that change is a gain or a loss (Kahneman and Tversky, 1979). Thus, the focus on total endowment in Figure 1 may be unrepresentative. Thirdly, if utility is proxied by money (through willingness to pay for instance), the criterion becomes problematic if utility and money are not perfectly correlated. There are a number of reasons why different people might have different valuations of money such as having differing baseline monetary resources, or simply valuing the ability to purchase differently. Layard (1972) discusses whether this second issue, while likely to be correct, should be accounted for in public decision making. He identifies a sequential approach in which total output is maximised (through the Kaldor-Hicks criterion), and then redistributed to move toward a more equitable distribution. However, he notes that this sequential approach is theoretically unsound,

“Can the size of the cake be maximised independently of who gets what? The answer would be yes, if transfers between people can be made without affecting their incentives to produce output. But unfortunately all practicable forms of transfer have some incentive effects: an obvious example is the ‘excess burden’ of the income tax.” (pp.57-58)

The Utility Principle, Individual Sovereignty and Consequentialism

I now move to the other characteristics of welfare economics which are parts of both the classical and neo-classical approach. The utility principle appears the least controversial at first glance. It seems difficult to consider a situation in which options are ranked in terms of utility, and the individual does not select the one that maximises that utility. However, there is scope for highlighting flaws in the principle relating to the use of the term ‘utility’. Dolan and Kahneman (2008) discuss the two leading interpretations of the term,

“The word ‘utility’ has two distinct meanings: it can refer either to the hedonic experience of an outcome or to the preference or desire for that outcome. These have been labelled experienced utility and decision utility, respectively.” (p.215)

Classical economics relied on experienced utility, see for example Bentham (1789). However, this idea was abandoned in the early twentieth century, with economists

becoming more reliant on decision utility (Fisher, 1918). Considering utility in terms of decision utility makes the utility principle tautological (and hence empty). This is because the decision is assumed to identify the option which has the highest utility. Considering it in terms of hedonic experience allows an individual to incorrectly order options (defined by decision utility or some other criterion), but leaves it difficult to judge as there is no straightforward way of measuring experienced utility.

The second element of welfare economics which might be questioned is the concept of individual sovereignty. This is closely associated with the utility principle. If individuals do maximise some function, but this function is flawed in some sense, it is debatable whether the preferred allocation of resources should follow the maximisation preferences of these individuals. A good example might be a situation in which an individual makes a decision based on limited evidence, and would likely make a different decision if this information deficit was removed. Another example might be a situation in which the individual has the information but continues to make a wrong decision (however that might be defined), such as the attitude of children towards school attendance. If individual sovereignty is compromised, so is the use of willingness-to-pay in a normative context (Rice, 1992; Rice, 1998). Approaches which reject individual sovereignty are however open to accusation of paternalism.

A third reason why welfare economics may need amendment is that consequentialism is a potentially contentious proposition. The view that the ends justifies the means may well be in conflict with the preferences of many people, in many different situations. The classical example of a refutation of consequentialism comes from Kant (1785 (translation 1959)), who argues that there are a set of categorical imperatives that are intrinsically valid. He defined morality as “*Always act(ing) according to that maxim whose universality as a law you can at the same time will*” (p.421). This issue is reflected in a health context in the growing literature base concerned with process utility (Birch and Donaldson, 2003; Tsuchiya, et al., 2005).

Cost-Benefit Analysis

Despite the difficulties associated with proxying markets, economic evaluation has become an important source of information for policy makers. The approach to

economic evaluation which has the longest history is cost-benefit analysis (CBA)³. Layard (1972) describes the basic notion of this,

“If we have to decide whether to do A or not, the rule is: Do A if the benefits exceed those of the next best alternative course of action, and not otherwise.”(p.9)

He redefines the benefits of the next best alternative course of action to A as the cost of A (as it is no longer done), to redefine the rule as,

“Do A if its benefits exceed its costs, and not otherwise” (p.9)

This reflects the Kaldor-Hicks criterion discussed previously. CBA is not widely used in healthcare economic evaluation. The major reason for this is the profound difficulty in valuing everything using a common metric. To state that benefits exceed costs or otherwise, we have to place a monetary value on health outcomes, something which have proven highly difficult. Rather, the increasingly dominant approach in the area is cost-effectiveness analysis (CEA). This differs from cost-benefit analysis in that it does not attempt to value all outcomes in a common unit⁴. Both CBA and CEA involve the comparison of two or more interventions according to their relative costs and outcomes (Drummond, et al., 2004). Indeed, some authors have argued that they are almost equivalent (Bala, et al., 2002; Donaldson, 1998; Phelps and Mushlin, 1991). Both CBA and CEA aim to assimilate evidence concerning cost, effectiveness and risk (insofar as it impacts on average effectiveness and cost), providing results designed to aid decision makers allocate resources appropriately.

Extra-Welfarism

To this point, I have identified the reasons why a *laissez-faire* approach to the healthcare sector may not be appropriate, and then looked at some reasons why a welfarist approach to evaluating health outcomes may be deficient. However, if we reject the *laissez-faire* approach and decide that a welfarist approach is inadequate, it is necessary to specify a framework within which health outcomes can be valued.

³ The methods used in this are in keeping with the Kaldor-Hicks criterion and welfarism more generally (these terms will be defined and investigated in the next section).

⁴ It might be argued that, even though CEA does not require valuation of outcomes in monetary terms, interpretation of the incremental cost-effectiveness ratio requires exactly that. In other words, CEA might merely defer the valuation of health outcomes.

The most likely candidate, and the one which has gained most traction, is extra-welfarism. Extra-welfarism is a widely-used alternative to welfarism in the evaluation of health interventions. Tsuchiya and Williams (2001) describe the relationship between welfarism and extra-welfarism in the following way:

“It is said that there are two ‘competing views’ on economic evaluation in health care. One is often seen as the ‘theoretically correct’ approach, that is based more firmly within the theory of welfare economics, whilst the other by comparison as some practical but not well formulated collection of rules of thumb (p.22)”

This comparison between the ‘*theoretically correct*’ welfarist approach and the ‘*practical but not well formulated*’ extra-welfarist one is appropriate. Any lack of formulation and coherence in extra-welfarism is a consequence of its numerous roots, its reliance on ideas from outside of economics, and the short period of time in which the approach has been developed relative to welfarism.

Respecifying the desideratum

Extra-welfarism emerged as a counterpoint to the perceived weaknesses in the assumptions underpinning welfarism and welfare economics. As far back as 1963, Feldstein argued

“... should not health care be allocated to maximise the level of health of the nation instead of the satisfaction which consumers derive as they use health services?” (Feldstein, 1963)

Decoupling certain items from utility is not a concept unique to health. James Tobin (1970) argued for specific egalitarianism, which argues that societal inequality aversion will differ from domain to domain,

“This is the view that certain specific scarce commodities should be distributed less unequally than the ability to pay for them. Candidates for such sentiments include basic necessities of life, health, and citizenship.”(p.263)

He then illustrates how the orthodox economic perspective differs, and foreshadows the issue of paternalism as a possible criticism of extra-welfarism that runs through this section,

“While concerned laymen who observe people with shabby housing or too little to eat instinctively want to provide them with decent housing and adequate food, economists instinctively want to provide them with more cash income. Then they can buy the housing and food if they want to, and, if they choose not to, the presumption is that they have a better use for the money.”(pp.263-264)

The decoupling of health and utility suggested by Feldstein and echoed in Tobin’s specific egalitarianism has fundamental implications for how a society makes decisions. It might be argued that the outcomes from a healthcare system should be in terms of health, and not be reliant on a monetary numeraire to value health outcomes. However, the less palatable consequence of this is that, under certain circumstances, it is permissible to ignore certain conventional (i.e. in terms of utility) potential Pareto improvements. If the person receiving the compensation would be willing to receive \$X to accept not receiving a health gain (and someone is willing to pay \$X to receive the health gain), the extra-welfarist position would be to argue that this is only a beneficial trade-off if the same Pareto-relativity applies in terms of health outcomes as well as money.

The roots of Extra-Welfarism

Coast (2009) identifies the various roots of extra-welfarism. These are the argument that basic goods should be allocated in a fair way (if the market cannot do so); the capabilities approach of Sen; government rejection of willingness to pay as the numeraire of benefit in allocating resources; the increasing role of decision makers in producing sources of values in public decision making; and the reality that many health economists were already analysing healthcare resource allocation decisions using health as the major outcome, albeit without a fully developed underlying principle for doing so. A good example of this last root is Williams’ seminal study of the cost-effectiveness of coronary artery bypass grafting which was important in placing economic concerns into national decision making processes (and provides a framework for doing so) (Williams, 1985).

Of these roots, the one that is not self-explanatory is Sen's capabilities approach. Hall *et al.* (2006b) describe Sen's movement away from welfare economics in terms of four key areas. Firstly, there is the decoupling of personal utility and individual choice; that is, that the individual can select an option which does not maximise his / her utility (which contradicts the utility principle discussed previously in the context of welfarism). Secondly, there is a criticism of utilitarianism, consisting of welfarism, sum-ranking and consequentialism. Thirdly, Sen (1992) introduces aspects of well-being beyond utility, namely functioning and capabilities. Regarding functionings, he says that

“Living may be seen as consisting of a set of inter-related ‘functionings’, consisting of beings and doings... The relevant functionings can vary from such elementary things as being adequately nourished, being in good health, avoiding escapable morbidity and premature mortality etc., to more complex achievements such as being happy, having self-respect, taking part in the life of the community and so on” (p.39)

Capabilities differ in that they reflect what the person might be able to achieve; even though someone may not climb Everest, having the opportunity to do so reflects positively on their well-being.

“In assessing our lives, we have reason to be interested not only in the kind of lives we manage to lead, but also in the freedom that we actually have to choose between different styles and ways of living. Indeed, the freedom to determine the nature of our lives is one of the valued aspects of living that we have reason to treasure.” (Sen, 2009)(p.227)

The fourth movement away from welfare economics, which is linked to the decoupling of individual utility and choice is that the individual preference for social states is affected by the individual's view of a social good. Sen (1992) illustrates this with an example,

“If a person aims at say, the independence of her country, or the prosperity of her community, or some such general goal, her agency achievement would involve evaluation of states of affairs in the light of

those objects, and not merely in the light of the extent to which those achievements would contribute to her own well-being.” (p.56)

Extra-welfarism reflects some of Sen’s work, particularly the emphasis on characteristics of the individual. However, as will be described, the practical implementation of extra-welfarism reintroduces sum-ranking and consequentialism albeit with outcomes other than utility being maximised. Brouwer (2008) defines extra-welfarism in terms of four differences in its approach from that taken under a welfarist framework:

“(i) it permits the use of outcomes other than utility;

(ii) it permits the use of sources of valuations other than the affected individuals;

(iii) it permits the weighting of outcomes (whether utility of other) according to principles that need not be preference-based; and

(iv) it permits interpersonal comparisons of well-being in a variety of dimensions, thus enabling movement beyond Paretian economics” (p.330)

Echoing Feldstein, the standard extra-welfarist approach places health itself as the central outcome (Culyer, 1991). Culyer argues that characteristics of people are important, including non-utility characteristics. If a characteristic of a person is that they have a need for healthcare (rather than a demand), the extra-welfarist framework would imply this supports providing an intervention for that person. The approach to outcome measurement that is most commonly taken in an extra-welfarist framework is to use life years or quality-adjusted life years (QALYs) gained. The strength of these outcome measures is that they are applicable across most areas of medicine. The assumptions underpinning this approach are discussed later in this chapter, and the methods for adjusting for quality are the central topic in Chapter 5. At this stage, it is important to note that, within the QALY model, outcomes are of value independent of who they accrue to, and also independent of the preferences of the individual receiving them.

Weinstein and Manning (1997) attempt to explain why extra-welfarism has become a major trend in outcome measurement in economic evaluation:

“Extra-welfarists, and many decision-makers in the real world of health care, are more willing to accept an approach that considers outcomes equitably (as CEA (Cost-Effectiveness Analysis) using QALYs does), rather than to accept an approach in which choices are heavily influenced by ability to pay” (p.127)

While the extra-welfarist framework gives an intuitive and conceptually straightforward solution for interpersonal comparison, it is unclear whether cost-effectiveness analysis using QALYs can easily consider outcomes equitably. While willingness to pay has considerable drawbacks in terms of unequal distribution of health outcomes, equality of the value of outcomes across individuals does not necessarily coincide with the concept of equity. This is a major issue; if equity is defined other than in terms of equality of gain in outcome, a QALY approach does not address equity issues without additional analysis such as an equity weighting system in which the value of outcomes is not independent of the person receiving the outcome. The description of how such a system might work, and an attempt to operationalise this, will be a major concern in subsequent chapters.

The focus now moves to how extra-welfarism has been implemented in practice, and the additional assumptions that have been made in this operationalisation. This is important as it teases out the additional assumptions and controversies that, while not necessarily intrinsic to an extra-welfarist framework, do exist in this dominant approach.

The Quality-Adjusted Life Year – A history and critique

The leading extra-welfarist approach to outcome measurement in the economic evaluation of healthcare is the quality-adjusted life year (QALY). One QALY is defined as one year of full health for one individual. Equally, a QALY can be generated through 6 months of full health for two individuals or two years of ‘half-health’ for an individual⁵. More generally, the number of QALYs resulting from a health profile is simply the product of the number of people, the number of years and a measure of quality of life such that full health and death are anchored at 1 and 0

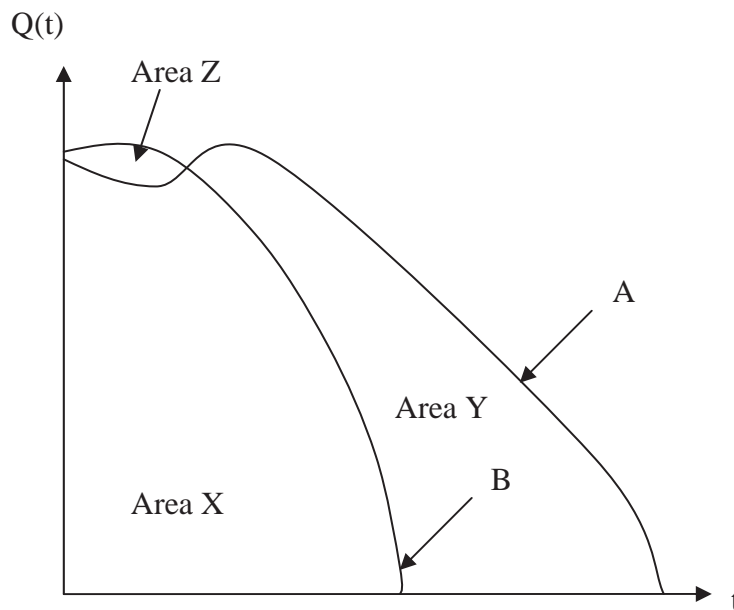
⁵ The issue of how to define health as (for example) ‘half-health’ is of course a major issue in the operationalisation of the QALY model. It has a very large literature base investigating the methods for doing so. Some existing conventions are discussed in Chapter 2, and an alternative approach attempting to remedy some of the existing deficiencies is presented in Chapter 5.

respectively. Using this approach, the impact of an intervention can be evaluated through a comparison of the number of QALYs produced with and without that intervention occurring.

QALY maximisation is essentially an extra-welfarist approach as it places health as the central focus of healthcare. Additionally, it allows interpersonal comparison, and their valuation results from more than just the preferences of the individual receiving them. As the main outcome measure in cost-utility analysis (CUA), it is the preferred measure in the major Health Technology Assessment processes worldwide, for example in Australia and England and Wales (Department of Health and Ageing, 2007; National Institute for Health and Clinical Excellence, 2007). Gold *et al.* (1996) identify that QALYs were developed in the late 1960s by researchers working across economics, operations research and psychology (Fanshel and Bush, 1970; Klarman, *et al.*, 1968; Packer, 1968). Further commonly cited work in the development of QALYs, and the movement from the operations research literature to the that of healthcare evaluation include studies by Torrance *et al.* (1972) and Williams (1985).

The intuitive appeal of the QALY is that it combines mortality and morbidity effects into one metric, allowing comparison between interventions in many areas of healthcare, both those which prevent death and those which reduce or remove the impact of chronic conditions. If we define health-related quality of life over time as $Q(t)$, this can be plotted in Figure 2.

Figure 2: Alternative Health Profiles over Time



Note that, in this case, health varies over time. This means that defining QALYs as the product of length of life, quality of life, and number of people experiencing the profile is not correct as quality of life differs over time. However, this simply means that the QALY produced under each program is simply the number of people experiencing the health profile multiplied by the area under the curve.

If two competing health programs A and B are considered in terms of the average health-related quality of life experienced by individuals receiving them⁶, the value placed on them by the QALY model is simply

$$TotalQALY = \int Q(t)dt. \quad \text{Equation 1}$$

Thus, the options A and B can be contrasted in outcome terms simply by comparing the difference between these areas. In Figure 2, Area X is common to both A and B. A receives Area Y, while B receives Area Z. Therefore, the incremental gain of A relative to B is $(X+Y) - (X+Z) = Y-Z$.

Gold argues that the QALY was initially derived from the theoretical underpinnings of welfare economics and expected utility theory (Pliskin, et al., 1980). The aim of economic evaluation is to maximise an outcome given scarce financial resources. However, welfarism would define this outcome as utility while modern CUA focuses

⁶ This assumes there is no time preference. This is not a significant assumption as it can be relaxed by assuming the $Q(t)$ term is adjusted using some discount rate.

on health as the outcome (which is an oddity given the CUA label). While health and utility will be strongly positively correlated, there remains significant justification for arguing that the modern use of the QALY in CUA has (for better or worse) moved away from its original foundations.

Evaluating the QALY

Loomes and McKenzie (1989) present three questions which require consideration when assessing the use of the QALY as the outcome resulting from healthcare interventions,

“(1) Whether any of the existing methods for eliciting quality of life valuations are reliable and valid

(2) Whether individual valuations can be scaled and somehow aggregated to give measures which enable legitimate interpersonal comparisons to be made

(3) Whether indeed the values to be used in social decision making should be some aggregate of individuals’ valuations” (p.304)

Regarding the first point, the quality of life valuation techniques in 1989 were rudimentary, and a considerable literature base has been developed in the area since then. However, each of the approaches for eliciting quality of life valuations has issues, and there is at yet no consensus on which approach can be best considered ‘reliable and valid’. This will be discussed at length in the next chapter. The idea of interpersonal comparison in point (2) depends on the validity of the extra-welfarist approach (or alternatively a classical welfarism which is not advocated in the health literature). Point (3) echoes Arrow’s Impossibility Theorem, which argues that basing social choices on individual values is impossible, assuming non-dictatorship (Arrow, 1950). A full discussion of this is not presented here: for a good discussion, see Deaton and Muellbauer (1980). The issue of whether we should, and whether we can use individual preferences in constructing social preferences is contentious, as summarised by Scanlon (1975),

“The fact that someone would be willing to forego a decent diet in order to build a monument to his god does not mean that his claim on others for aid

in his project has the same strength as a claim for aid in obtaining enough to eat...” (pp.659-660)

Loomes and McKenzie raise three very important considerations in how we value health outcomes for economic evaluation, and these ideas will be returned to throughout the thesis. However, I think it is important to identify two other areas in which the QALY model may be deficient, each of which will be discussed now. Firstly, it is arguable that there are circumstances in which maximising of health is not the major goal of the health sector. An example of this in a specific condition (cystic fibrosis) will be outlined. Secondly, a considerable literature has developed regarding the constraints the QALY model places on individual preferences. This will be discussed subsequently.

Using the QALY – A problematic example

It is worthwhile to consider whether this simple QALY model can truly capture our conception of gain resulting from an intervention in all circumstances. Radhakrishnan *et al.* (2008) considered the existing evidence regarding the cost-effectiveness of screening approaches for cystic fibrosis (CF). A brief consideration of the issues surrounding use of the QALY model in this case identifies some potentially important issues, some relevant to economic evaluation in a subset of health areas relating to reproductive health, but some relevant more generally.

Briefly, cystic fibrosis (CF) is the most common life-shortening genetic disease, with an incidence of 1 in 2500 (Welsh, et al., 2001). Average life expectancy is only in the mid to late 30s (Cystic Fibrosis Australia, 2009). There is still no cure, the daily therapies are rigorous and there are many years of ill health before death.

Screening for CF carrier status in parents is becoming increasingly common. If both parents are identified as carriers of a CF gene, any resultant infant has a 25% chance of having CF. Under the QALY approach, the benefits of screening would be limited to the health improvements elicited by screening. There are two serious concerns in this context. Judging the cost-effectiveness of any screening program in terms of health alone means that the true negative results are of no value. Issues such as reassurance (which a true negative would provide) have been identified as important outcomes for evaluating health services (Haas, 2005). Equally, the distress caused by a false positive result is ignored if health is the only outcome. The issue of outcomes

of value not being captured within the QALY model is one which is by no means limited to reproductive health.

In terms of issues more specific to reproductive health, this disease area identifies questions which the QALY model finds difficult to answer. It is common that potential parents identified as both being CF carriers abstain from future reproduction. How does this fit into the QALY model? Do we assume that the QALYs that would have been accrued by the CF baby in its life are lost? Do we assume that the parents undertake IVF and have a ‘replacement baby’? These options are very unpalatable, and indicate that the use of the QALY model in this context leads to both uncertainty regarding the appropriate calculation, but also consequences that were not intended by those who initially developed the technique.

Individual preference constraints in the construction of the QALY model

Having noted some limits of the QALY model through the cystic fibrosis example, I now turn to the constraints the QALY model places on individual constraints. Viney and Savage (2006) outlined a simple model of individual health care decision making, upon which the QALY model can be constructed given certain additional constraints. Individual utility over a lifetime is assumed to be dependent on health h and consumption c of other non-health related goods over T periods of time,

$$U = U((h_1, c_1), \dots, (h_T, c_T)). \quad \text{Equation 2}$$

When considering two options, for example a treatment and no treatment, the individual is uncertain regarding the likely levels of health and consumption of other goods taking either option would entail. Therefore, the value of a profile of health P , is a function of health and consumption payoffs of all m possible outcomes, plus the various probabilities p_i assigned to each

$$P = [((H_1, C_1); p_1), \dots, ((H_m, C_m); p_m)], \quad \text{Equation 3}$$

and

$$(H_i, C_i) = (h_{i1}, c_{i1}), \dots, (h_{iT}, c_{iT}). \quad \text{Equation 4}$$

The individual then selects the option which maximises the total utility associated with any option. This general framework encompasses the QALY model but the QALY model imposes a number of restrictions on individual preferences. Pliskin,

Shepard and Weinstein (1980) presented a set of constraints that the QALY model places on individual preferences, a set which was refined by Bleichrodt, Wakker and Johannesson (1997). Thus, I will introduce the former work, but note that the more relevant set of constraints comes from the latter.

Pliskin, Shepard and Weinstein (1980) identified three conditions that have to be imposed on individual preferences over health gambles to ensure it can be described by the QALY model. These are utility independence, constant proportional trade-off and risk neutrality on life years.

Utility independence is defined by Pliskin *et al.* (1980) in the following way,

“Let Y and Z denote two attributes of the outcome of concern (e.g. Y = life years, Z = health status). Attribute Y is utility independent of attribute Z if preferences for lotteries over Y, with Z held at a fixed level z_0 , do not depend on the particular level z_0 . Attributes Y and Z are mutually utility independent if Y is utility independent of Z and Z is utility independent of Y.” (p.208)

If health state A for 5 years is preferred to health state B for 5 years, utility independence asserts that A is preferred to B for any period of time. Is this plausible? Clearly, there is a very strong relationship between the two attributes Y and Z. Firstly, if we consider longevity, it is usual that y_0 would be preferred to y_1 if it is relatively longer. However, if the health state were so poor that it was considered worse than immediate death, the ordering of the preference would switch, and violate utility independence. If it were possible to categorise health states z_n to be better or worse than death, then utility independence could be amended to reflect this special case. However, this is not necessarily a categorisation that can be made. This is because it is plausible that Z is not utility independent of Y (i.e. the reverse utility independence). Consider two similar health states z_1 and z_2 . Both have considerable limitation in possible activity. The difference between them is that z_1 is a health state that can be adapted to, while z_2 is a health state which remains equally poor as Y increases. For short values of Y, it is plausible that z_1 is the worse state, but that the relativity is reversed for longer Y. Having established that the relative utility of two health states can cross zero as Y increases, it is apparent that a health state can be preferred to death at some values of Y, but not at others. Therefore, the adaptation to redescribe Y

as being utility independent of Z is not possible. Bleichrodt and Johannesson (1997) tested utility independence and found it was not supported.

Constant proportional trade-off is defined in the following way by Pliskin *et al.* (1980),

“The constant proportional trade-off assumption of life years for health status is said to hold if the proportion of remaining life years that one is willing to give up for an improvement in health status from any given level q_1 to any other level q_2 does not depend on the absolute number of remaining life years involved.”(p.210)

Thus, if I had five years to live, and was willing to sacrifice one of these years (i.e. 20% of the remainder) to gain full health for those four years, constant proportional trade-off would necessitate that, if I had ten years to live, I would be willing to sacrifice two years (i.e 20%) to return to full health. Does this correspond with our preferences for different lengths of life? I would argue it is plausible other than at extreme levels of remaining life years. If I had one month to live, I am not sure that I would be willing to sacrifice 20% of that to gain full health. However, whether the average person would be willing to sacrifice more or less than 20% is uncertain. It is plausible that the average respondent would not be willing to sacrifice any of such a small endowment of life expectancy (or would be willing to sacrifice a much smaller percentage). However, it is plausible that a person in this situation would not be willing to endure any ill-health, so would be willing to sacrifice more than 20%. Regarding empirical evidence, Bleichrodt and Johannesson (1997) tested for constant proportional trade-off and found more support for it than for utility independence.

Bleichrodt, Wakker and Johannesson (1997) identify that fewer assumptions are required to impose the QALY model on preferences. They show that it requires only that risk neutrality for life years in every state be imposed and that the zero condition, in which a life expectancy of zero has a utility of zero (irrespective of quality of life) is asserted. Since the latter is surely correct, the validity of QALYs depends on the issue of risk neutrality. They begin by stating that a lottery be defined as $[p_1, (Q_1, T_1); \dots; p_n, (Q_n, T_n)]$ with the individual receiving an outcome (Q_i, T_i) with probability p_i . This satisfies the von Neumann-Morgenstern axioms (von Neumann and Morgenstern, 1947) and there is a utility function U such that the utility of the

gamble is equal to $(p_1 U(Q_1, T_1) ; \dots ; p_n U(Q_n, T_n))$. In choosing between sets of lotteries, it is this composite value which determines choice.

They then define risk neutrality as being

“(when)...quality of life (is) held fixed, the individual is indifferent between a lottery over life years and the expected life duration of that lottery.”(p.109)

This means that $U(Q_1, T_1)$ is linear with respect to Q_1 and T_1 . Since one can add a constant to a utility function without changing the utility function, the zero condition (which is widely acknowledged to be uncontroversial) can be imposed on the utility function by adding minus the constant to anchor everything to pass through zero. This leads to Theorem 1 in Bleichrodt *et al.* (1997), which asserts,

“Under expected utility, the following two statements are equivalent for a preference relation on lotteries over chronic health states:

(1)The QALY model holds:

$$U(Q, T) = V(Q)T$$

(2)The zero-condition holds and, for each health state, risk neutrality holds for life years. Q.E.D.”(p.110)

The evidence regarding risk neutrality for life years suggests it is unrealistic (McNeil, et al., 1978; Stiggelbout, et al., 1994), an unsurprising result given the previously cited evidence regarding the assumptions required by Pliskin *et al.* (because, as noted by Bleichrodt, Wakker and Johannesson (1997), a test of mutual utility independence or constant proportional trade-off is implicitly a test of risk neutrality for life years).

What then should we conclude about the impact of these issues on the appropriateness of the QALY-model? This is a difficult question; while the restrictions it places on preferences are suspect, and it is based on ethical premises which are not universally agreed upon, it may still represent the best approximation of the value of health gains, and it has been shown to be operationalisable. The exploration of non-linearity of utility with respect to time is considered at length in Chapters 5 and 6. If it is concluded that a non-linear utility function is a better representation of preferences than a linear one, the next question is whether we can plausibly include such a pattern in healthcare decision making.

Some additional criticisms of extra-welfarism

Three additional criticisms of extra-welfarism can be made. Firstly, there are many issues surrounding the description of health and the methods for valuing a health state. The problems will be described in detail in the next chapter, with a potential solution presented in Chapter 3, and tested in Chapter 5.

The second criticism of extra-welfarism which should be noted here is that of paternalism. Paternalism is a possible criticism of extra-welfarism as it removes judgment regarding the value of an outcome from the person receiving that outcome. This is unavoidable once the choice has been made to move the focus of analysis from utility to health.

The third criticism of extra-welfarism refers to the decision making rules within an extra-welfarist framework (Coast, 2009). Coast argues that the rule of maximisation, which is acceptable in a welfarist framework, becomes untenable in an extra-welfarist framework. In a welfarist framework, efficiency and equity are separate, with utility maximisation between individuals ensuring efficiency and post-intervention redistribution aiming to meet equity concerns. However, in an extra-welfarist framework, health is the outcome; compensation cannot occur through redistribution. As Coast argues,

“The production and distribution of health in the extra-welfarist paradigm is not theoretically separable – a greater total quantity of health cannot be produced and then reallocated around the public by dint of taxes and subsidies and so to redistribute health to satisfy requirements of fairness is not physically possible.”(p.789)

There are two obvious responses to this argument, one of which is flawed, and the other reliant on a difficult assumption. The first response is to say that extra-welfarism need not take such a narrow view of outcome measurement. Rather, it can consist of health and any number of other things, meaning that compensation to achieve an equitable distribution of outcomes can take place in domains other than health. Indeed, Culyer (1991) argues that extra-welfarism supplements traditional welfarism, potentially including any number of additional aspects. However, as Coast notes, this argument is flawed as practical applications of extra-welfarism have all reduced down to measuring outcome through health and health alone (Coast, 2009), with the

possible exception of rule of rescue-type situations (McKie and Richardson, 2003). The second response is to argue that simple maximisation, which is a product of potential Pareto improvement in this area, can exist even if compensation cannot occur. In a society in which many decisions regarding health allocation are made, it is possible that simple maximisation of health in every decision might leave all members of society better off and thus not requiring compensation. This is clearly a brave assumption, and unlikely to be defensible in most circumstances.

Beyond Welfarism and Extra-Welfarism

To this point, I have defined the debate regarding outcome measurement in health-based economic evaluation as one between welfarists and extra-welfarists. While these are the two most likely candidates for frameworks in which normative judgements can be reached, the debate would be deficient if it ignores alternative standpoints which use only some (or none) of the tenets of either.

In this section, a number of criticisms and alternative approaches to the concept of economic evaluation will be considered. Other than extra-welfarism, no alternative framework to welfarism has gained traction in terms of actual societal decision making. However, the alternatives and arguments outlined below suggest a number of ways in which the current approaches are lacking, and may hint at future research directions.

Communitarianism

A leading example of an alternative approach to outcome measurement in health settings comes from Mooney (2009; 1998; 2005), who discusses the merits of communitarianism. He identifies four main strands of thought which contribute to this idea. Firstly, to identify the ideas on which decision making takes places in the health sector, it is the view of the community that matters rather than those of anointed experts. This appears to contrast with the paternalism which is a likely consequence of an extra-welfarist approach (Brouwer, et al., 2008); however, the avoidance of paternalism is an issue communitarianism has not unequivocally addressed, an issue which will be discussed below. The second strand of communitarianism is that community views are not necessarily the aggregation of individual preferences (a view upon which welfarism is based, and is followed to some degree within extra-welfarism). Thirdly, the social good associated with health has to be established.

Finally, interpersonal comparability is not possible, particularly if health and / or health need are different constructs in different social groupings.

Communitarianism shares with extra-welfarism the belief that the source of utility should not be limited to the consumption of goods and services. It also argues that the individual is making decisions in the context of the broader society and hence that the image of the individual being a rational self-interested being choosing freely between alternatives is incorrect. This leads to individuals selecting options which impact negatively on their own utility if the option brings about an adequately large societal benefit (Mooney and Russell, 2003).

Mooney (2009) argues that the focus on the society providing the framework for decision making is much more than the consideration of externalities that can be easily built into either a welfarist or extra-welfarist framework. Rather, he argues that a constitution is required under which the '*modus operandi, the culture, and the governance of the organisation* (is described)' (p.109)

How this emphasis on the community per se rather than a group of individuals might impact on outcome measurement in economic evaluation is unclear. Mooney argues that "*(t)he upkeep of the health of the population is a community obligation*" (p.127), and that acceptance of this obligation is an important factor in the social determinant of health.

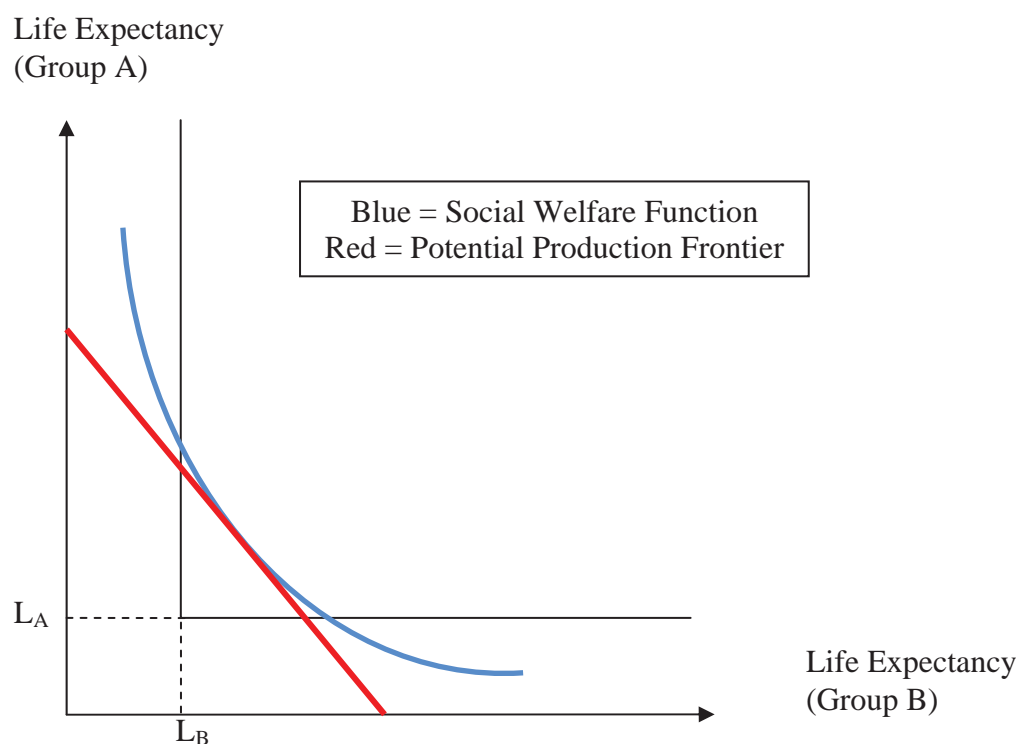
One possible criticism of communitarianism is that it is unclear how it differs from moral relativism. Moral relativism argues that the moral correctness of something is always contingent on the society in which it takes place; therefore, there is no such thing as absolute right and wrong. This is a contentious view, and leads to unpalatable conclusions.

Empirical Ethics

Empirical ethics presents a different view to either communitarianism or extra-welfarism regarding how social preferences should be discerned. It agrees with the idea that social welfare is not merely the sum of individual utilities; however, it considers that we can begin to capture social welfare through surveys coupled with the provision of appropriate information to help inform the decision.

Where the values derived from empirical ethics differ from those generated as population values is in the explicit and central consideration of ethics. Richardson and McKie argue that resources should be allocated in accordance with ethically justified population values (Richardson, 2002a; Richardson and McKie, 2005), something which Hall *et al.* (2006b) identify as a constrained maximisation problem. In many (probably most, arguably all) situations, this constraint will not be required. Imagine a Social Welfare Function considering life expectancy in two groups A and B such as the blue line given in Figure 3 that is generated through some society-wide survey.

Figure 3: An Unconstrained Social Welfare Function



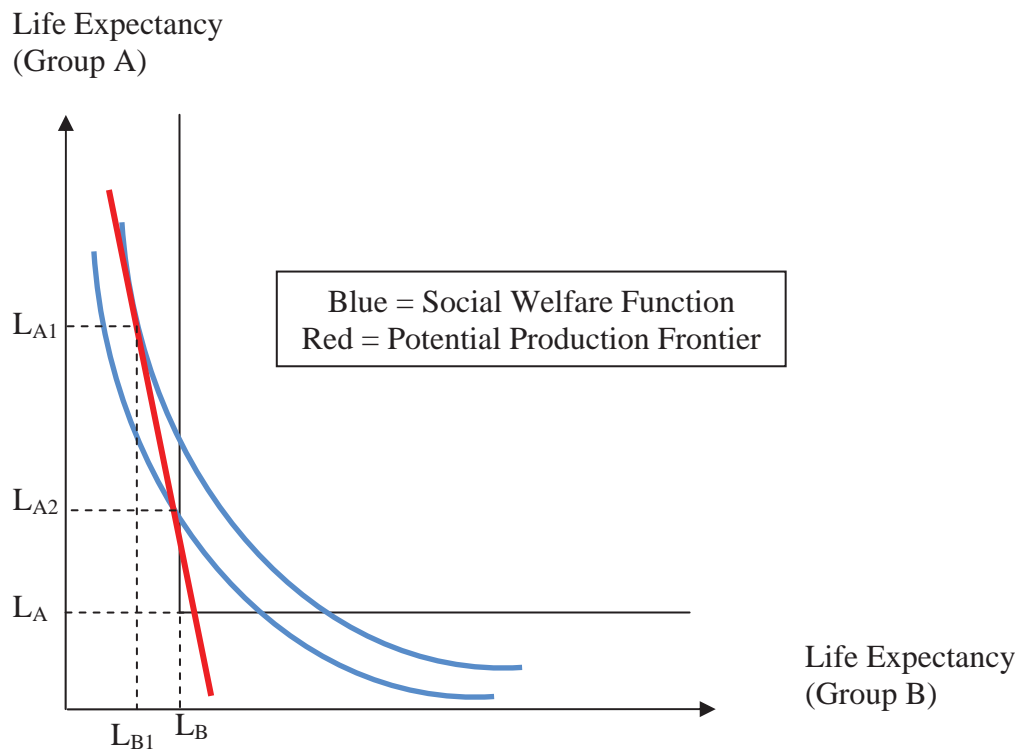
At each point on the Social Welfare Function, social welfare is equal; therefore, the convexity of it shows the society is averse to inequality. This can be seen because, if the life expectancy of one group is relatively low, societal welfare can be kept equal by simultaneously removing a relatively large amount of life expectancy from the group with longer life expectancy and giving a relatively small amount of life expectancy to the group with the shorter life expectancy.

Imagine now a potential production frontier in red. This gives the combination of life expectancies in the two groups that are possible. Finally, imagine an ethically justified

constraint that says that all individuals have the right to L years of life, marked for each group as L_A and L_B . Thus any position to the left of L_B , or below L_A is ethically unacceptable. In this case, the constraint plays no role, as the highest Social Welfare Function can be reached within the L-shaped area in the top right hand corner which reflects the ethically justified combinations of life expectancies in the two groups.

Consider the same situation, but with a different potential production frontier.

Figure 4: A Constrained Social Welfare Function



In this case, the slope of the potential production frontier is much steeper. The maximum social welfare is found where individuals in Group A receive L_{A1} and individuals in Group B receive L_{B1} . However, since $L_{B1} < L_B$, this position is considered unethical. Therefore, the optimum point is where Group A receives L_{A2} and Group B receives the minimum acceptable life expectancy L_B . While this is optimal under Empirical Ethics, there is a clear loss associated with constraining life expectancy in this way (as the solution is on a lower Social Welfare Function).

While theoretically clear, this constraint is difficult to operationalise. Firstly, there is the issue of how surveys can be used to elicit accurately a Social Welfare Function. This is considered in Chapter 6 on Equity Weights for Economic Evaluation.

However, at this point it is important to note that the formulation of the question can significantly impact on inferences made. The second and probably more pressing issue is that it is very difficult to identify a set of ethical constraints that are unquestioned. While many would agree with Richardson that “*(e)thical analyses should be of pivotal importance in establishing the normative foundation of policy analysis*” (Richardson, 2002b), the form of these analyses will differ and produce very different constraints. Clearly, they cannot rely on investigation of what people find acceptable or otherwise (or the result would never constrain the conventional social welfare maximising position). Therefore, an *a priori* approach is required but who should determine the basis of this analysis is unclear. If this is not determined, this uncertainty seems to make the approach untenable as a practical alternative.

To this point, a number of ideas have been discussed. The role of economic evaluation as a proxy for a market where a market does not, can not, or should not exist has been described. Then, the framework within which outcome measures are based has been outlined, with the main conflict being between welfarists and extra-welfarists.

Alternatives to welfarism were then presented. While both communitarianism and empirical ethics relax assumptions that many would feel to be unhelpful, they each have problems associated with implementation that may have contributed to their lack of traction. On the other hand extra-welfarism has become a central part of decision making in countries that undertake economic evaluation of new health technologies (Department of Health and Ageing, 2007; National Institute for Health and Clinical Excellence, 2007). Implementation is not necessarily reflective of the acceptability of the assumptions underpinning the extra-welfarist approach; it is likely to be as much related to the relatively easy applicability and transparency of the approach.

Initial conclusions

The lessons to take from this chapter concern the uncertainty that typifies outcome measurement in economic evaluation in health, even before we consider the instruments used to describe health or the valuation techniques used to place these on some index. Arrow’s work has been discussed, showing that a purely market driven solution is likely to lead to often unpalatable consequences. If we as a society decide to intervene and make decisions about how resources flow between groups and individuals, we face a variety of new and possibly equally intractable issues. We do

not know what domains should be considered when evaluating outcomes. If we are willing to make a decision on this, we do not know what we should objectively try to do once we have identified this appropriate domain. Finally, the dominant approach, namely extra-welfarism and the use of a QALY-type outcome, makes simplifying assumptions to allow societal decision making to be tractable. These assumptions are not particularly defensible, and may lead to disagreement between government actions and societal perspectives. Whether the approach towards maximisation in societal decision making should reflect that taken by a typical individual within it is uncertain, but the divergence between the two is important and the decision to overrule how individuals make decisions should at least be an explicit, and preferably justified.

Thesis structure

The investigation of these issues is structured as follows. Chapter 2 considers some major approaches to outcome measurement in the economic evaluation of health interventions. This will include the symbiotic relationship between quality of life instruments and preference elicitation approach, and will aim to answer Loomes and McKenzie's (1989) concern that the methods for eliciting quality of life valuations be '*reliable and valid*'. Chapter 3 will present the discrete choice experiment (DCE) as a tool for exploring complicated preference patterns. This includes methods for analysing the complementary dimensions, tools for deriving welfare measures from DCE results, and techniques for exploring response heterogeneity. Chapter 4 will introduce the notion of efficient DCE design, and describe the techniques which are employed in this thesis. Chapter 5 will present a discrete choice experiment using the methods outlined in Chapters 3 and 4 to answer some of the questions posed in Chapter 2. Thus, using a discrete choice experiment, utility weights for the SF-6D will be estimated. A secondary output from this chapter will be that it will investigate the constraints assumed by the QALY model regarding individual preferences. Chapter 6 will investigate the assumption contained within the QALY model that the objective of healthcare decision making is to maximise the total health of the population. This will be done through equity weights for economic evaluation, using a similar methodology to that introduced in Chapters 3 and 4, and tested in Chapter 5. This potentially allows the relaxation of the assumption that a health gain has value independent of who it accrues to. Chapter 7 will summarise the findings of the previous chapters and attempt to identify the areas in which outcome measurement is

performing well and poorly, and potential opportunities for allowing a more representative and fit-for-purpose approach.

Chapter 2: Measuring health-related quality of life – standard and novel approaches

Chapter summary

In this chapter, the focus turns to how health states within the descriptive system are assigned a value, and then how health-related quality of life is described for the purposes of outcome measurement in health economic evaluation. The chapter begins by investigating the preference elicitation techniques used for valuing health states (e.g. Standard Gamble, Time Trade-Off), including a discussion of the strengths and weaknesses of each approach. The chapter then looks at the various generic multi-attribute utility instruments (which will be termed as MAUIs) available to analysts. These instruments are potentially highly valuable because they aim to describe health in a generic way, allowing comparability across disease areas. The relative merits of each are considered, and conclusions are reached regarding whether it is possible to identify one or more as preferred. This exploration of MAUIs is intended to be instructive for the data collection and analysis that follows in Chapter 5. The imputation of values for those health states not directly valued is then addressed; this issue becomes increasingly important as the number of health states directly valued as a proportion of the total within an instrument falls. The purpose of describing methods for valuing health states, both directly and through imputation is to identify some limitations of the existing approaches, which might be addressed by other preference elicitation techniques such as discrete choice experiments. The methods for using these as an alternative to the current approaches are described in Chapters 3 and 4, and then tested in the subsequent empirical chapters.

Introduction

The measurement of health-related quality of life for the construction of the QALY involves placing an individual with a particular combination of health-related characteristics on to a scale with full health valued at one, and death valued at zero. The reason for this constraint should be clear from Figure 2 in the introductory chapter. It allows for states to be worse than death, consequentially with a value less than zero, so the zero value is an anchor rather than a boundary. To undertake these valuations, both a generic multi-attribute utility instrument for describing health and a method for scoring specific health states within that instrument are required.

This chapter is divided into three sections. Section A discusses the merits of the various approaches to scoring individual health states. Section B outlines the merits and weaknesses of generic quality of life measurement relative to disease specific measurement, and outlines the major instruments that have been used to describe health. The reason for doing this is that Chapter 5 will attempt to value health states within one of these, and it is important to determine if one can be identified as preferable, or in some sense, more appropriate. Section C then considers how the values of states which are not directly valued are imputed; this is of particular importance in instruments with large numbers of possible health states.

Section A: Methods for valuing health states

The three most common approaches to scoring health states are the Standard Gamble, Visual Analogue Scale and the Time Trade-Off. For examples of each, see Brazier *et al.* (2002), Devlin *et al.* (2003) and Dolan (1997) respectively. Each will be outlined now, with a discussion of the relative merits of each. The Person Trade-Off has also been advocated in the area (Nord, et al., 1999), but has not been widely used to estimate utility weights for economic evaluation, and hence is not discussed here. A possible alternative to these (the discrete choice experiment) will be described in Chapters 3 and 4, and tested in Chapter 5. It is the Standard Gamble that is replaced in the empirical work presented in Chapter 5; therefore, I will start by describing this approach.

Standard Gamble

The Standard Gamble is based on von Neumann-Morgenstern Expected Utility (von Neumann and Morgenstern, 1944), under which the value of a health profile P under uncertainty to be

$$P = [((H_1, C_1); p_1), \dots, ((H_m, C_m); p_m)], \quad \text{Equation 5}$$

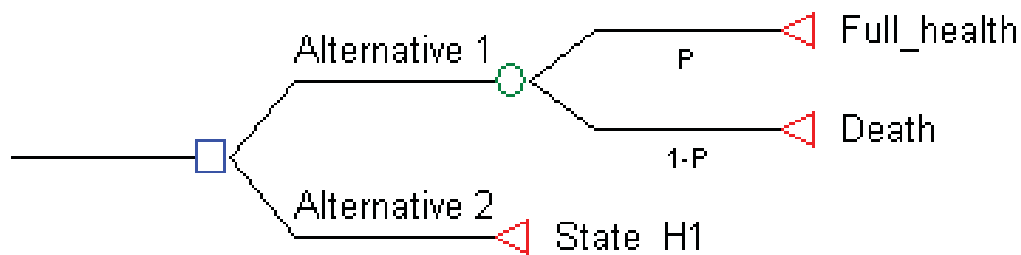
where the value P is dependent on the mutually exhaustive set of probabilities p of attaining a level of health H and a level of other consumption C . In a conventional Standard Gamble, it is assumed the C terms are constant, leaving P to be determined simply by a set of health profiles over time and a corresponding set of probabilities.

At its most generic, the Standard Gamble asks the respondent to choose between

- a certain outcome
- a gamble with one outcome better, and one worse than the certain outcome (with probabilities that sum to 1), and the aim is to find the probability at which these two options are equivalent.

A common approach in the Standard Gamble methodology is to ask respondents to choose between the certain prospect of an imperfect health state, and a gamble with a probability of receiving full health and a complementary probability of immediate death. Option A is therefore health state H_I for time T for certain, while option B is a gamble with probability p of full health for time T , and a probability $(1-p)$ of immediate death. This is displayed graphically in Figure 5.

Figure 5: The Standard Gamble



Assuming that $v(.)$ is the value of a health profile, by adjusting p until the respondent is indifferent between options A and B, the value of H_I can be found:

$$p \cdot v(FH) + 0 = v(H_I) \quad \text{Equation 6}$$

In this way, the value of full health is anchored at 1 (as the individual would not accept any risk of death in the gamble if the certain option was full health), and the value of death is anchored at 0 (as the individual would always take the gamble if the certain prospect was death).

It is not necessarily true that the poor option in the gamble is death. In situations in which health state H_I is relatively mild, the value of p at which the respondent would be indifferent between the two alternatives would be very high. Indeed, it is likely that the majority of people would accept no risk of death to avoid H_I . In these cases, it is common for the death option to be replaced with a poor health state. The advantage of this approach over using death in the Standard Gamble is that a wider range of p would be observed. However, to do this, it is necessary to be able to place a value on

the poor health state that has replaced death; this is sometimes done through a chaining approach in which a poor health is valued relative to immediate death, and then all other health states are valued relative to this poor health state. For an example of this, see Brazier (2002).

There are a number of criticisms that can be applied to the Standard Gamble methodology. Firstly, it has been shown that individuals find extreme probabilities (less than 0.1 or greater than 0.9) difficult to use (Kahneman and Tversky, 1982).

A second criticism of the way this approach has been applied in health is that it has assumed risk neutrality of the respondent. While the ordinality of scores for health states is likely to be correct under this method, the valuation of health states is increasingly overestimated relative to true valuation as the respondent's risk aversion increases⁷. This pattern was identified by Torrance (1976) and has been demonstrated consistently. In the extreme risk-averse case, the respondent would be unwilling to accept any risk of death to gain full health. Strictly speaking they would never reach a point of indifference as an increase in p below 1 would not entice them to switch their preference away from the certain prospect, while a p of 1 would not be a point of indifference as the certain prospect of the imperfect health state would be weakly dominated by the alternative.

The relevance of individual attitudes to risk and uncertainty for societal decision making is a difficult issue, and one which was discussed by Arrow and Lind (1970), who stated that,

“(I)n private capital markets, investors do not choose investments to maximise the present value of expected returns, but to maximise the present value of returns properly adjusted for risk. The issue is whether it is appropriate to discount public investments in the same way as private investments” (p.364)

Arrow and Lind go on to outline some positions on this question. I will not expand on these here, other than to conclude that the question is contentious and those who argue that individual attitudes to things like risk and uncertainty should not be reflected in

⁷ See Figure 10 for evidence of a Standard Gamble-derived valuation system producing relatively high scores relative to one derived using a Time Trade-Off

societal decision making would consider the use of Standard Gamble methodology to be flawed.

Time Trade-Off

The Time Trade-Off approach uses a similar framework as that identified by Standard Gamble, only identifying strength of preference for a health state by adjusting duration rather than risk. It was first identified as a potentially useful tool for the valuation of health states by Torrance (1976). Rather than assigning probabilities to different health profiles and asking the respondent to consider a gamble, the P term consists of one possible outcome (rather than m), and adjusts the level of H and critically also the time over which the individual experiences H .

For a health state H_1 preferable to immediate death, the trade-off is typically between 10 years in this health state and x years in full health (at which quality of life is assumed to be one). The value of a health profile has been defined in the introductory chapter as

$$TotalQALY = \int Q(t)dt \quad \text{Equation 7}$$

The aim of the Time Trade-Off is to identify a position at which

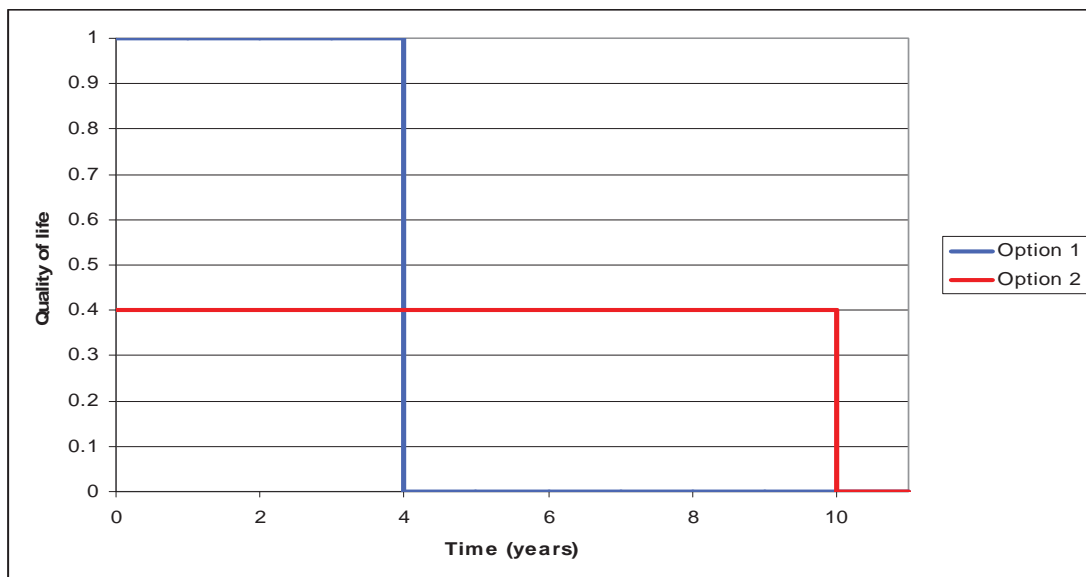
$$TotalQALY_{H_1,10years} = TotalQALY_{Fullhealth,xyears} \quad \text{Equation 8}$$

As the health states in the Time Trade-Off are constant over time,

$$\int_1^{10} H_1 dt = \int_1^x 1 dt \quad \text{Equation 9}$$

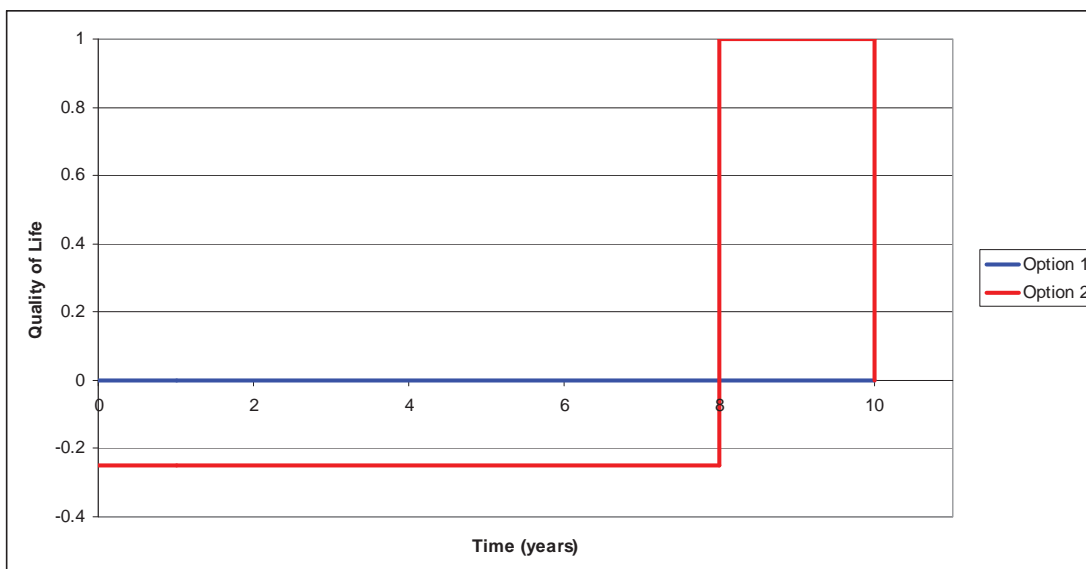
At the point of indifference, this implies that $10(H_1) = x$. As x defines that point, a value can be placed on H_1 . Graphically, this is illustrated in Figure 6.

Figure 6: The Time Trade-Off for states considered better than death



For states considered worse than death, the approach would collapse as there would be no point of indifference between the two prospects. The task is typically changed to a choice between the certain prospect of immediate death and a combination of x years in H_I , followed by $(10-x)$ years in full health, followed by death. At the point of indifference, the value the individual places on the $(10-x)$ years in full health exactly counterbalances the negative value placed on the x years in the poor health state. This is illustrated in Figure 7.

Figure 7: The Time Trade-Off for states considered worse than death



Once this point of indifference has been found, it is necessary to estimate a utility weight based on this result. One approach is to define the utility weight as $(-x / (10 - x))$. The issue with this approach is that the lower bound is likely to fall at a very low value. If the respondent is indifferent when $x=9.5$, the weight would be -19 . The impact of assuming such a weight for a health state is that it would have a significant (and probably disproportionate) impact on results of economic evaluations. One option is to take this value, and apply a linear transformation to it, such that the lowest value is set at -1 (Tilling, et al., 2010). The problem with this solution is two-fold. Firstly, it is sensitive to the maximum value of x (Norman, et al., 2009). Secondly, the -1 to 1 scale remains fundamentally arbitrary. Assigning a value to a health state better than death has a clear interpretation. For example, if a health state is valued at 0.5 , it means a respondent is willing to sacrifice half of their life expectancy to return to full health (or in the Standard Gamble context, to accept a 50% chance of death) A value of, for example, -0.5 has no comparable interpretation. Nevertheless, it is accepted that economic evaluation has to allow for states to be worse than death, and a solution has to be identified. One solution applied in an Australian study is to value the health state as $(x/10)-1$ (Viney, et al., 2011b). Again, this has the arbitrary constraint that the minimum possible score is negative one.

Overall, the valuation of health states worse than dead is very difficult, and no adequate solution to the issue has yet become standard. A promising possibility is the use of the Lead Time TTO (LT-TTO), proposed by Robinson and Spencer (2006), and tested by Devlin *et al.* (2011). In this approach, each profile in a TTO is preceded by a 'lead time' in full health. Then, the study aims to find the point at which a respondent is indifferent between 1) the lead time (say 10 years) in full health, followed by 10 years in the health state being valued; and 2) x years in full health. If a health state is better than dead, x will lie between 10 and 20 years. If a health state is worse than dead, x will lie below 10 years. Of course, it is plausible that the respondent would trade off all of the lead time; in this case, the LT-TTO might have to adjust the times involved in the experiment, such as by extending the lead time, or changing the ratio of the lead time to the health state being valued. This is a promising alternative TTO approach, but is yet to be standard practice.

Problems with the conventional Time Trade-off have been identified by a number of sources. The first issue is that the task is cognitively challenging, particularly in

situations where the health state is worse than immediate death. Uncertainty regarding comprehension of the task leads to predictable problems with health state valuations. The process by which the point of indifference is reached involves bouncing between increasingly less extreme values of x until the individual is indifferent. This move towards a position where $x=5$ continues until the respondent:

- is indifferent between x years of full health and 10 years in H_I
- (if x is less than 5) prefers x years in full health to 10 years in H_I
- (if x is greater than 5) prefers 10 years in H_I to x years in full health

When faced with a choice in a Time Trade-Off, the respondent will have three options, two of which will cause the bouncing between extreme values to cease. If the respondent is uncertain about the task (or acting strategically to end the task promptly), the likelihood of reaching a point of indifference at which x is close to 5 is minimal (Norman, et al., 2010).

Two other issues arise regarding the use of TTO. Firstly, time preference is a major concern in the interpretability of TTO-derived quality of life scores (Attema and Brouwer, 2010; Norman and Viney, 2008). The framework shown diagrammatically in Figure 7 assumes that there is no time preference; as time preference is introduced into individual preferences, the value of ten years in the state being valued decreases more than the value of the period in full health and consequently, the value assigned to the health state is artificially deflated. While Attema and Brouwer (2010) argued that it is possible to adjust TTO scores for time preference, this requires the analyst to specify a universal discount rate, which is arguably no more realistic than assuming individuals do not discount future health benefits.

Finally, in the health state valuation literature, the period in H_I has typically been fixed at ten years. This has allowed comparability between responses, but the applicability of results generated using this value is dependent on the linearity of utility with respect to duration. Dolan (1996) explored the issue, and concluded that this assumption needs to be treated with caution.

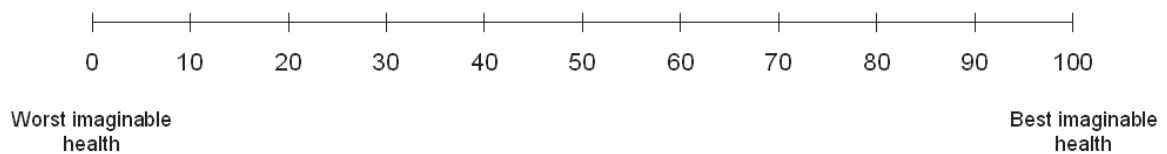
A criticism of the TTO relative to the Standard Gamble is that the latter is more grounded in the Expected Utility Theory of von Neumann and Morgenstern (1944). Therefore, Standard Gamble is often seen as the gold standard in the area (Gafni,

1994). However, it is notable that both SG and TTO impose the same conditions on the utility function in order to generate utility weights (risk neutrality, constant time preferences).

Visual Analogue Scales

Visual Analogue Scales (also known as category rating scales) are lines with defined endpoints where respondents can place items. They emerged from the psychophysics literature where “*there was an interest in measuring people’s perceptions of various objective phenomena, such as heat and sound.*”(p.84) (Brazier, et al., 2007). A simple example of a health Visual Analogue Scale is given in Figure 8.

Figure 8: Health Visual Analogue Scale



The approach to valuation in the Visual Analogue Scale is to ask the respondent to place a hypothetical health state on the scale, illustrating the quality of life in the health state relative to full health and death. Interval properties are assumed; therefore, an improvement from 0 to 20 is of the same magnitude as an improvement from 60 to 80. The idea of using these scales for the measurement of health benefit is long-standing (Patrick, et al., 1973); indeed, they have been investigated alongside a range of generic quality of life instruments (Brooks, et al., 2003; Feeny, et al., 2002; Sintonen, 1994). In the example given in Figure 8, it would be necessary to consider that some health states are worse than being dead. Therefore, Brazier *et al.* (2007) suggest that, for economic evaluation, death be placed on this scale, and scores be rescaled such that death is valued at zero, and full health is valued at one (p.84).

The VAS has a significant advantage over TTO and SG of being simple to complete, requiring only one question per state rather than an iterative process. Ordinal preferences collected through a VAS process are likely to be valid also. The final advantage is that VAS scores are not impacted by risk attitudes or time preference, which are of concern in the SG and TTO respectively.

However, there are substantial problems associated with assuming a cardinal property of data collected using a VAS. The most pressing criticism is that VAS-derived scores are prone to end-of-scale bias or central tendency bias (Patrick and Erickson, 1993; Streiner and Norman, 1995). Finally, while Standard Gamble is based on von Neumann-Morgenstern preferences (von Neumann and Morgenstern, 1944), and involve a notion of sacrifice in valuing a health state, there is no easy description for what VAS-derived health scores actually mean. The QALY model is based on the sacrifice that people are willing to make when trading off health-related quality of life and length of life; this aspect of sacrifice is not in the VAS and represents a major limitation of the approach.

Section A Conclusion

The three dominant approaches to valuation of individual health states have been outlined, as have some issues regarding the reliability of the valuations derived from them. Visual Analogue Scales are the simplest of the three, but subject to considerable issues relating to the interpretability of responses. Time Trade-Off and Standard Gamble both have some attractive features. Both conform to Expected Utility Theory space, although make restrictive assumptions about time preference and risk attitude respectively. Since time preference and risk attitude are likely to differ both over respondents, and situations faced by one respondent, it is not possible to adjust for these factors. The patterns of responses for the Time Trade-Off suggested it was highly sensitive to the way the question was asked, and the schema the analyst uses to reach a point of indifference. This is likely to apply to the Standard Gamble also. An alternative approach to both will be considered in Chapter 5.

Section B: Multi-Attribute Utility Instruments

The methods described in Section A can in principle be applied to any health state. The focus of this chapter however is on a specific subset of the instruments used to describe health, namely generic multi-attribute utility instruments (MAUIs). These are the most commonly applied quality of life instruments in the economic evaluation of healthcare. Briefly, these instruments are defined by two characteristics:

1. They describe health in a way which is not specific to particular health conditions; thus, they aim to allow comparison between changing health profiles in different disease areas.

2. They aim to identify the relative importance of different aspects of the health they describe. In other words, they allow the analyst to infer acceptable trade-offs between both different aspects of ill-health, and between aspects of ill-health and survival.

I will begin this section by discussing the merits of MAUIs relative to other quality of life instruments. I will then discuss who should be responsible for valuing health states, before introducing Torrance's framework within which MAUIs can be evaluated. Finally, I will look at the various MAUIs that can be used in the construction of the QALY.

Categorising approaches to describing and valuing quality of life profiles

The appropriate approach for describing and valuing health-related quality of life is dependent on the motives the analyst has for doing so. There are two major choices which will be explained and discussed here. The first choice for the analyst is whether to describe health using disease-specific vignettes or using a more generic approach. The distinction is that the former describes health in dimensions considered most relevant to a particular field of investigation (for example aspects of vision in the management of patients with macular degeneration), while the latter attempts to describe health in a way which is applicable across multiple disease areas (this distinction will become clearer in the investigation of generic MAUIs discussed later). Both approaches are clearly of value. Disease-specific measurement is most likely to be of use if the analyst wishes to test hypotheses regarding whether the quality of life of an individual is likely to change following an intervention (or whether it changes differently from the change experienced in some comparator group). If suitably constructed, it is more sensitive to changes in quality of life than a generic measure, because it emphasises aspects of quality of life that are most plausibly improved in the specific patient group undergoing a particular health intervention. If changes in quality of life in a patient group are more likely to be in terms of (for example) mental health, a disease-specific instrument can focus on these aspects to the exclusion of aspects of health-related quality of life that will not be impacted on by such an intervention.

The use of disease-specific measures is therefore limited in situations where benefits between disease areas require comparison, for example in economic evaluation in

which a decision maker has to allocate scarce resources between competing programs. For a generic tool to be of most value, it must capture areas of potential health gain which may be applicable under all branches of medicine, health care more generally, and health in the broadest sense. This leads to an inevitable conflict between the burden placed on the individual survey respondent in terms of number and complexity of the questions they must face and the ability of the tool to capture changes in all possible facets of health, an important issue, and one which will be returned to later.

Whose values?

A second issue arises when deciding whose opinions are of interest when a value has to be placed on a health profile. Two leading contenders are to use either the general population or a patient group to value particular health states. In the context of economic evaluation, the reason for doing the former is that the values assigned to health states are used for societal decision making, rather than as a prediction for individual decision making. Thus, surveying a general population to value health profiles intends to identify the value society places on changing levels on health. The advantage of using a group experiencing poor health to derive values for health states similar to those they are themselves experiencing is that their valuations are more likely to reflect the true disutility associated with the poor health state (Nord, et al., 1999). To some degree, the discussion of whose values should count reflects some of the issues raised in the introductory chapter concerning the move towards extra-welfarism. Moving away from the utility of the individual receiving the healthcare is a key part of extra-welfarism and therefore it is unsurprising that the dominant approach in the economic evaluation literature reflects that. Dolan (2011) discussed the issue of hedonic adaptation in this context. He argues that a general population sample is likely to overstate the disutility associated with a health state, an issue which has been identified elsewhere (Loewenstein and Angner, 2003). This is because focusing on valuing a health state causes the dimensions presented in the state to be considered to the exclusion of other parts of an individual's life. This phenomenon has been termed elsewhere as a focusing illusion (Schkade and Kahneman, 1998; Wilson and Gilbert, 2003). A further explanation for divergence between patients' and general population values for health states is that the general population might underestimate the possibility of adapting to manage a loss of health (McTaggart-Cowan, et al., 2011). As I will be investigating the valuation of health for use in economic evaluation, my

fieldwork in Chapter 5 will focus on a general population sample. However, this does not imply that the use of patient values is without merit.

A framework for building and evaluating MAUIs

Before describing the various MAUIs that might be used in construction of the QALY, it is necessary to look at the ways in which these health states are constructed (and how they can be evaluated). Torrance (1986) described a framework within which generic quality of life instruments (of which multi-attribute utility instruments are a sub-set) exist. He discusses both the depth and breadth of coverage as being dimensions in which the descriptive ability of a generic quality of life instrument can be assessed. By breadth of coverage, he means the range of domains (or dimensions) that health can manifest. These might include physical function, sensory function, pain, cognitive ability and so on. The term 'depth of coverage' refers to the detail in which each domain is described by the levels of the dimension. An example Torrance provides divides physical functioning into mobility, physical activity, self care and role performing, and, as an example of further sub-division, divides self-care into dressing, bathing, continence and eating. An individual health state is defined as any combination of levels of the various domains and sub-domains.

It is tautological that, assuming extra domains are areas in which health manifests itself, these additional domains increase the descriptive ability of an instrument. It is comparable to adding parameters to an Ordinary Least Squares regression and comparing the unadjusted R^2 value. Equally, finer sub-divisions of domains and sub-domains do likewise. However, there are balancing forces that mean this extra descriptive ability is not cost-free. For example, placing a utility value on each of these possible states is an increasingly difficult task as either the breadth or the depth of the instrument increases. It is within this context that generic quality of life instruments have been developed. On the one hand, they aim to capture as many domains of health as possible, and in as much detail as is needed to identify any significant health effect stemming from an intervention. On the other, adding additional dimensions and / or levels causes a dramatic increase in the number of possible health states. Thus, the analyst is forced to consider increasingly sophisticated techniques to generate utility values for each of the possible individual health states, or to employ an increasingly large sample. A central theme of this

chapter is that the damaging impact of this increased reliance on imputation and assumption has to be balanced against the benefits of increased depth and breadth of the instrument.

Structural independence

Before describing and evaluating the multi-attribute utility instruments, it is worthwhile considering one possible characteristic of these instruments which is of value for an analyst. Hawthorne *et al.* (2001) provide a number of these characteristics that make a generic multi-attribute utility measure suitable for use in economic evaluation. One important focus is on the concept of structural independence. They describe a situation where this is met as one in which a single attribute should not be measured in more than one way (von Winterfeldt and Edwards, 1986). Feeny *et al.* (2002) take a less restrictive view of structural independence, defining it to be that

“it is plausible and logically possible for a person at a particular level in one attribute to be at any level in each of the remaining attributes.”
(p.114)

Feeny’s view is less restrictive as it implies that some correlation is acceptable, but a necessary relationship between levels of different attributes is not (while von Winterfeldt and Edwards’ description rejects the acceptability of both). Whichever attitude is followed, this is an issue in each of the instruments described below as there is likely to be considerable correlation between dimensions of a particular individual. Indeed, some correlation is inevitable between dimensions of a generic multi-attribute utility instrument; exploring and minimising it is important, but the requirement of structural independence has to be balanced against the sensitivity of the instrument.

Conversely, there may also be combinations of levels of dimensions that cannot plausibly co-exist for an individual. Hawthorne *et al.* (2001) argue that there is a natural tension between structural independence and sensitivity, presumably as extra depth in the descriptive instrument can lead to dimensions being highly correlated. This is a convincing argument as each additional dimension has less health space to describe and is therefore more likely to overlap with pre-existing dimensions. They argue the solution to this is to establish statistical independent through methods such

as factor analysis. This is described in more detail later in the context of Brazier’s construction of the SF-6D through factor analysis of the SF-36 (Brazier, et al., 2002).

The chapter will now describe the key generic multi-attribute utility instruments, and discuss some of the strengths and weaknesses of each.

EuroQoL - 5 Dimensions (EQ-5D)

The EQ-5D was designed by the Euroqol group, a team of researchers covering Europe, North America, Africa, Australasia and Asia. In its most well-known iteration, it consists of five dimensions each with three possible levels (therefore $3^5 = 243$ individual health states). The English-language version of this iteration is reproduced in Table 1.

Table 1: The EQ-5D

Dimension	Level	
Mobility	1	I have no problems in walking about
	2	I have some problems in walking about
	3	I am confined to bed
Self Care	1	I have no problems with self-care
	2	I have some problems washing and dressing myself
	3	I am unable to wash and dress myself
Usual Activities	1	I have no problems with performing my usual activities
	2	I have some problems with performing my usual activities
	3	I am unable to perform my usual activities
Pain / Discomfort	1	I have no pain or discomfort
	2	I have moderate pain or discomfort
	3	I have extreme pain or discomfort
Anxiety / Depression	1	I am not anxious or depressed
	2	I am moderately anxious or depressed
	3	I am extremely anxious or depressed

Before looking at criticisms of the EQ-5D, it is useful to consider the nature of the disutilities considered by the instrument. I would argue that there are two distinct types of disutility considered. In the Mobility, Self-Care and Usual Activities dimensions, the levels explicitly reflect the types of capabilities that Sen emphasises as important (1980; 1992). For the Pain / Discomfort and Anxiety / Depression dimensions, this is not true; rather the EQ-5D talks about health-related characteristics

of the individual. Clearly, the two types of disutility have some correlation in that someone with poor capability in a dimension is also likely to have poor health-related characteristics in that dimension also. Talking about health characteristics implies certain capabilities (e.g. If I imagine I have extreme pain or discomfort, I may infer the impact this has on my capabilities) and visa versa. However, this inference is not perfect, and the emphasis on either capabilities or health-related characteristics may impact on how the general population sample value health states with combinations of each. If the respondent tasked with valuing health states is more concerned with one type of disutility (capabilities or health descriptions), then the dimension which is explicitly couched in terms of that type of disutility may be a larger driver of an aggregate measure of quality of life.

Regarding structural independence, there is clearly scope within the EQ-5D for dimensions to be correlated within an individual. Additionally, there are combinations of levels that seem highly implausible, for example Mobility Level 3 (“I am confined to bed”) is highly unlikely to co-occur with Self-Care Level 1 (“I have no problems with self-care). A useful investigation of structural independence in the context of the EQ-5D is to consider the distribution of individuals across the five levels. This has two important components. Firstly, in a general population sample, are any levels largely unused? This is important as the sensitivity of the instrument would be negatively affected if the number of levels used in practice is reduced; this is clearly a potentially damaging issue for the EQ-5D which has only three levels in each dimension. Admittedly, these instruments are designed to be used for people with a degree of ill-health (who would form a relatively small proportion of the general population). However, having few respondents at a particular level suggests an instrument may not be sensitive to changes in interventions aimed at relatively healthy people. Secondly, self-assessed health can be used to investigate correlations between dimensions, and hence structural independence. Currently unpublished data from an Australian National Health and Medical Research Council Project Grant (403303) undertaken at the Centre for Health Economics Research and Evaluation, University of Technology, Sydney gathered self-assessed health data from an Australia-representative population (n=2,494). This population were members of an online panel with respondents stratified by age and gender to match Australian population distributions. Each respondent completed the EQ-5D (and the SF-6D, which will be

discussed later in this chapter) before undertaking a further task (some of which is discussed in Chapter 5 of this thesis). The simple breakdown of responses of the 2,494 individuals by dimension is given in Table 2.

Table 2: Self-Assessed Health (EQ-5D) (n=2,494)

Number (%) of responses	Level 1	Level 2	Level 3
Mobility (MO)	1,920 (77.0%)	572 (22.9%)	2 (0.1%)
Self-Care (SC)	2,377 (95.3%)	112 (4.5%)	5 (0.2%)
Usual Activities (UA)	1,956 (78.4%)	514 (20.6%)	24 (1.0%)
Pain / Discomfort (PD)	1,361 (54.6%)	1,037 (41.6%)	96 (3.9%)
Anxiety / Depression (AD)	1,563 (62.7%)	832 (33.4%)	99 (4.0%)

Level 3 is very rarely used for Mobility and Self-Care, and to a lesser extent Usual Activities. This may be partially driven by the sample being a population one (and hence having few sick people). Nevertheless, the skewness in these dimensions is striking. For Self-Care, the skewness of results is most apparent with over 95% of respondents saying they have no problem in that dimension. The second issue, that of correlation between dimensions, can be investigated by assuming the levels to be continuous and calculating correlation coefficients. The results for this are presented in Table 3.

Table 3: Correlation coefficients between self-assessed EQ-5D dimensions

	MO	SC	UA	PD	AD
MO	1.000				
SC	0.387	1.000			
UA	0.622	0.432	1.000		
PD	0.550	0.324	0.543	1.000	
AD	0.219	0.184	0.311	0.306	1.000

There is considerable correlation between attributes, particularly Mobility and either Usual Activities or Pain/Discomfort, Usual Activities and either Self-Care or Pain/Discomfort. In dimensions with higher correlation coefficients, the issue of structural independence becomes increasingly problematic.

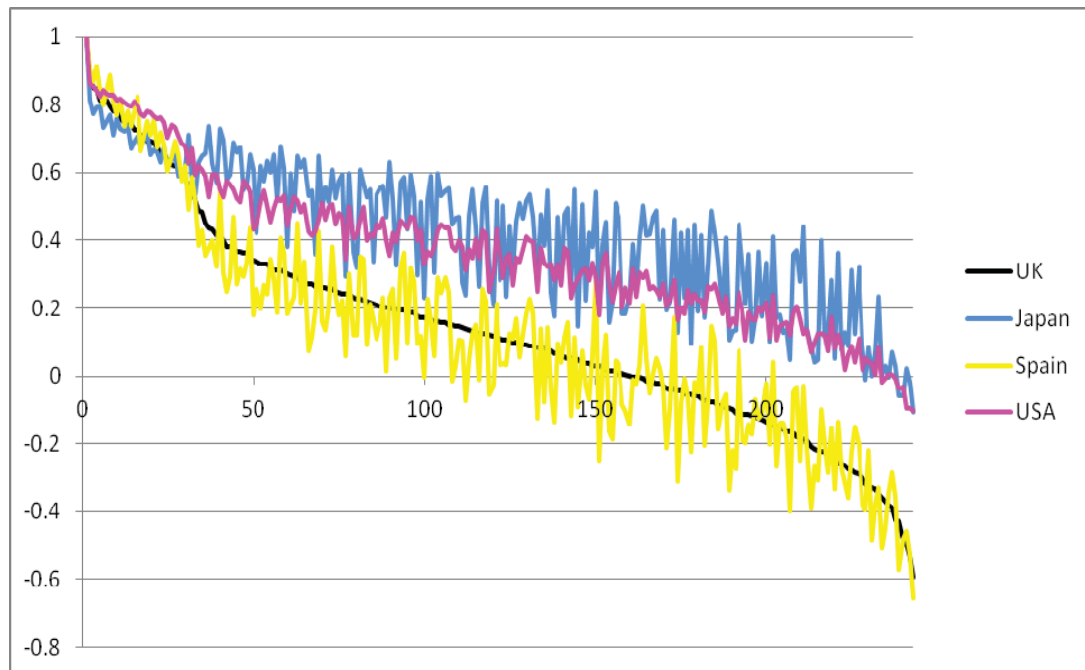
One potential strength of the EQ-5D is that it has been valued in a number of different countries, potentially allowing different jurisdictions to assess changes in quality of life using the societal attitudes to ill-health specific to their citizens. Table 4 is adapted from Szende *et al.* (2007) and demonstrates the countries with published algorithms for the EQ-5D.

Table 4: Existing EQ-5D Algorithms

Country and Reference	Sample Size	Valuation Method
Argentina (Augustovski, et al., 2009)	611	Time Trade-Off
Australia (Viney, et al., 2011b)	417	Time Trade-Off
Australia (Viney, et al., 2011a)	1,039	Discrete Choice Experiment
Belgium (Cleemput, 2003)	722	Visual Analogue Scale
Denmark (Wittrup-Jensen, et al., 2001)	1,686	Visual Analogue Scale
Denmark (Wittrup-Jensen, et al., 2001)	1,332	Time Trade-Off
Europe (Greiner, et al., 2003)	8,709	Visual Analogue Scale
Finland (Ohinmaa, et al., 1996)	1,634	Visual Analogue Scale
Germany (Claes, et al., 1999)	339	Visual Analogue Scale
Germany (Greiner, et al., 2005)	339	Time Trade-Off
Japan (Tsuchiya, et al., 2002)	621	Time Trade-Off
Netherlands (Lamers, et al., 2006b)	309	Time Trade-Off
New Zealand (Devlin, et al., 2003)	1,360	Visual Analogue Scale
Poland (Golicki, et al., 2010)	321	Time Trade-Off
Slovenia (Prevolnik Rupel and Rebolj, 2001)	733	Visual Analogue Scale
South Korea (Jo, et al., 2008)	500	Time Trade-Off
Spain (Badia, et al., 1997)	300	Visual Analogue Scale
Spain (Badia, et al., 2001)	1,000	Time Trade-Off
UK (MVH Group, 1995)	3,395	Visual Analogue Scale
UK (Dolan, 1997; MVH Group, 1995)	3,395	Time Trade-Off
USA (Shaw, et al., 2005)	4,048	Time Trade-Off
Zimbabwe (Jelsma, et al., 2003)	2,440	Time Trade-Off

However, this weight of evidence has certain limitations. While each algorithm provides values for the country in which the study was undertaken, it is unclear what other countries should do if they wish to use EQ-5D valuation algorithms. This is not a trivial problem in that countries with well-developed traditions in economic evaluation are absent from the list (such as Canada and, until recently, Australia). This represents a potential problem as the uncertainty allows considerable scope for strategic behaviour. Figure 9, which is a reproduction of a figure used by Norman *et al.*, (2009) illustrates the issue.

Figure 9: Comparing UK results with other leading studies



On the x-axis, the UK algorithm developed by Dolan (1997) has been used to rank the 243 EQ-5D states. The UK curve (in black) is therefore smooth as the valuations (on the y-axis) fall monotonically as the states are ranked by their valuation from highest to lowest. The other three lines show the valuation placed on the 243 states in other studies in Spain, the USA and Japan (Badia, et al., 2001; Shaw, et al., 2005; Tsuchiya, et al., 2002). The values placed on EQ-5D health states under the UK and Spanish algorithms are consistently below those from the other two countries, and consequentially have a considerably greater spread between the better health states and the poorer ones. Therefore, when considering a typical health state prior and subsequent to some intervention, the incremental change will tend to be larger in the UK and Spanish algorithms. Thus, an intervention which improves quality of life will show a lower cost per QALY if these algorithms are used: When ranking healthcare interventions, using the UK or Spanish algorithm will relatively favour interventions improving quality of life rather than life expectancy. Furthermore, even in situations in which countries do not show considerable difference in willingness to trade quantity of life for improved quality of life (such as Spain and the UK), the choice of algorithm can still differ considerably, as seen by the oscillation of the Spain algorithm around the UK values. High oscillation identifies that the relative importance of the five dimensions differs between that country and the base country.

The criticism of the EQ-5D in terms of poor sensitivity to small changes in health-related quality of life (HRQoL) is partially addressed by Janssen *et al.* (2008a; 2008b), and by Herdman *et al.* (2011) who expand the instrument to have five levels in each dimension. This was done by inserting an additional level between level 1 and level 2, and another between level 2 and level 3 (these relate to slight problems and severe problems respectively). The results to date are promising, suggesting good face validity and improved discriminatory power relative to the standard EQ-5D. However, this is likely to represent a different choice in the trade-off continuum between description and ease of valuation, rather than a fundamental step forward. Also, while it represents an improvement in terms of the depth of the instrument, it does not address breadth in that dimensions of quality of life not adequately addressed by the standard EQ-5D remain neglected under the adapted approach.

Short Form – 6 Dimensions (SF-6D)

The SF-6D is an adaptation of the SF-36 specifically for use in economic evaluation (Brazier, et al., 2002). The SF-36 is also a generic quality of life instrument, but not a multi-attribute utility instrument as it does not quantify the trade-offs between levels. It contains eight dimensions (vitality, physical functioning, bodily pain, general health perceptions, physical role functioning, emotional role functioning, social role functioning and mental health) (Ware, et al., 1993). In the physical functioning dimension for example, there are ten items each with three levels corresponding to '*limited a lot*', '*limited a little*' and '*not limited at all*'. These are coded as 1, 2 and 3, and summed to provide a score between 10 and 30. This is then rescaled on to a 0-100 scale. However, these scores are not appropriate for use in economic evaluation. They neither consider the trade-offs between different dimensions (so each dimension is implicitly equally important) nor place health states on a zero-to-one scale required for the construction of a QALY (although the latter point could be remedied with a simple transformation).

The SF-36 has been reduced in size using factor analysis, the consequence of which was the SF-12 (Ware, et al., 1995). This new tool identified which components of the SF-36 were most important in determining overall well-being, and produced a new instrument focusing on these. However, while this addresses the onerous nature of the SF-36, scores from the SF-12 are no better than those from the SF-36 in the context of

economic evaluation as they do not consider the relative importance of the levels of the chosen dimensions. Brazier and colleagues (2004; 2002) built on this work, excluding all general health questions and combining role limitation due to physical problems and role limitation due to emotional problems (but retaining the distinction in the labels attached to levels). These two studies produced the SF-6D, one derived from the SF-12 and one from the SF-36.

The six dimensions of the resulting SF-6D instrument are physical functioning (PF), role limitation (RL), social functioning (SF), pain (PA), mental health (MH) and vitality (VI). For the SF-6D derived from the SF-36 (which is the one used in this thesis), the dimensions have between four and six levels. Consequentially, the complete factorial contains $6^2 \times 5^3 \times 4 = 18,000$ health states. This SF-6D is reproduced below:

Table 5: The SF-6D

Dimension	Level	
Physical Functioning	1	Your health does not limit you in <i>vigorous activities</i>
	2	Your health limits you a little in <i>vigorous activities</i>
	3	Your health limits you a little in <i>moderate activities</i>
	4	Your health limits you a lot in <i>moderate activities</i>
	5	Your health limits you a little in <i>bathing and dressing</i>
	6	Your health limits you a lot in <i>bathing and dressing</i>
Role Limitation	1	You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems
	2	You are limited in the kind of work or other activities as a result of your physical health
	3	You accomplish less than you would like as a result of emotional problems
	4	You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems
Social Functioning	1	Your health limits your social activities <i>none of the time</i>
	2	Your health limits your social activities <i>a little of the time</i>
	3	Your health limits your social activities <i>some of the time</i>
	4	Your health limits your social activities <i>most of the time</i>
	5	Your health limits your social activities <i>all of the time</i>
Pain	1	You have <i>no pain</i>

	2	You have pain but it does not interfere with your normal work (both outside the home and housework)
	3	You have pain that interferes with your normal work (both outside the home and housework) <i>a little bit</i>
	4	You have pain that interferes with your normal work (both outside the home and housework) <i>moderately</i>
	5	You have pain that interferes with your normal work (both outside the home and housework) <i>quite a bit</i>
	6	You have pain that interferes with your normal work (both outside the home and housework) <i>extremely</i>
Mental Health	1	You feel tense or downhearted and low <i>none of the time</i>
	2	You feel tense or downhearted and low <i>a little of the time</i>
	3	You feel tense or downhearted and low <i>some of the time</i>
	4	You feel tense or downhearted and low <i>most of the time</i>
	5	You feel tense or downhearted and low <i>all of the time</i>
Vitality	1	You have a lot of energy <i>all of the time</i>
	2	You have a lot of energy <i>most of the time</i>
	3	You have a lot of energy <i>some of the time</i>
	4	You have a lot of energy <i>a little of the time</i>
	5	You have a lot of energy <i>none of the time</i>

As a result of its construction, there are a number of noteworthy points that should be made regarding the instrument. In certain dimensions, the quantifier refers to the proportion of time with an impediment to full functioning (e.g. “You have a lot of energy none of the time”), while in others, the quantifier is the degree of impediment (e.g. “You have pain that interferes with your normal work (both outside the home and housework) *extremely*”). This is potentially important as a person may treat the two quantifiers differently.

While the EQ-5D effectively imposes monotonicity (i.e. Level 2 in any dimension is worse than Level 1 in that dimension, and similarly for Level 3 relative to Level 2), this is not so strictly imposed in the SF-6D. A good example of this is in Levels 2, 3 and 4 of Role Limitation. Level 2 specifies the problem to be physical, while the others specify emotional problems. This was done to allow a reduction in the number of dimensions (Brazier, et al., 2002), but has to be acknowledged when the health states are valued in the sense that the disutility associated with being at Level 2 relative to being at Level 1 need not be smaller than that associated with being at Levels 3 or 4.

As with the EQ-5D, the SF-6D contains elements of both types of health disutility discussed previously, namely capabilities and health description. The first four dimensions covering physical functioning, role limitation, social functioning and pain are mostly based around capabilities, while mental health and vitality are not. The same criticism of inconsistency that was applied to the EQ-5D can be applied here.

As with the EQ-5D, it is potentially instructive to consider the patterns of responses for self-assessed health within the SF-6D to explore structural independence. As before, the 2,494 respondents who described their own health were tabulated, and had the following distribution and correlations across levels:

Table 6: SF-6D Self-Assessed Health (n=2,494)

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
PF	881 (35.3%)	909 (36.45%)	400 (16.0%)	255 (10.2%)	34 (1.4%)	15 (0.6%)
RL	1,587 (63.6%)	608 (24.4%)	137 (5.5%)	162 (6.5%)		
SF	1,443 (57.9%)	565 (22.7%)	321 (12.9%)	125 (5.0%)	40 (1.6%)	
PA	971 (38.9%)	762 (30.6%)	395 (15.8%)	171 (6.9%)	128 (5.1%)	67 (2.7%)
MH	727 (29.2%)	1,142 (46.8%)	447 (17.9%)	150 (6.0%)	28 (1.1%)	
VI	178 (7.1%)	1,062 (42.6%)	748 (30.0%)	426 (17.1%)	80 (3.2%)	

PF = Physical Functioning; RL = Role Limitation; SF = Social Functioning; PA = Pain; MH = Mental Health; VI = Vitality

Table 7: Correlation Coefficients between self-assessed SF-6D dimensions

	PF	RL	SF	PA	MH	VI
PF	1.000					
RL	0.572	1.000				
SF	0.673	0.675	1.000			
PA	0.663	0.595	0.679	1.000		
MH	0.375	0.526	0.510	0.422	1.000	
VI	0.582	0.500	0.548	0.519	0.512	1.000

PF = Physical Functioning; RL = Role Limitation; SF = Social Functioning; PA = Pain; MH = Mental Health; VI = Vitality

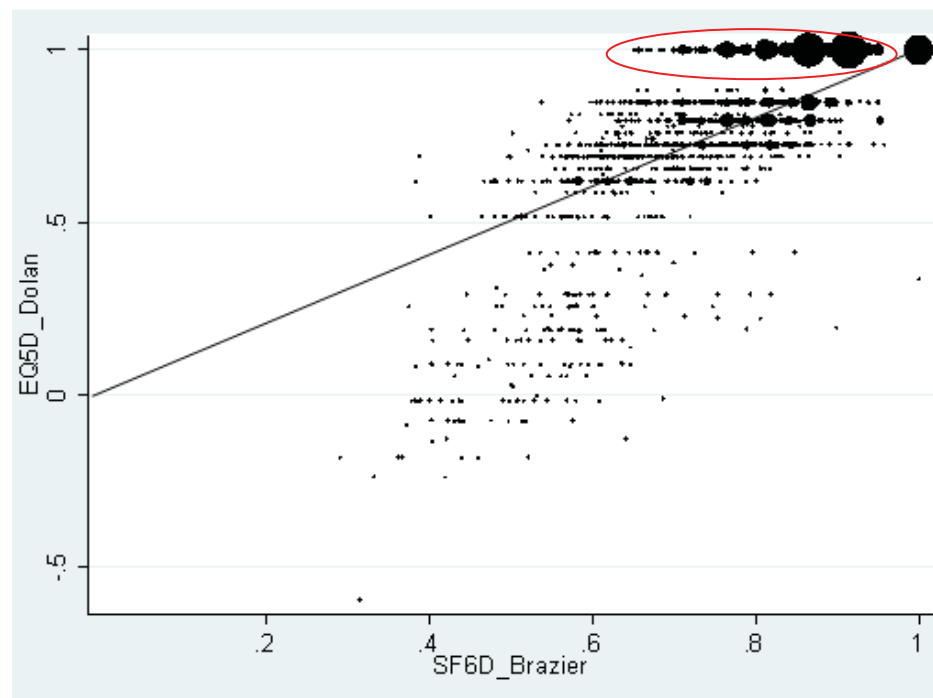
As with the EQ-5D, there is a tendency towards the better levels, which is to be expected given the sample and the aim of the instrument to cover all possible health

states. The degree to which self-assessed health clusters at Level 1 is smaller in the SF-6D. However, the correlation coefficients are high, as in the EQ-5D.

Bearing these issues in mind, what can we then say regarding the relative merits of the EQ-5D and the SF-6D? It is clear that the SF-6D has greater breadth and depth of the areas of health-related quality of life. However, as has been argued, the choosing of extra breadth and depth poses issues in the valuation of individual health states.

However, it is noteworthy that the SF-6D allows for smaller decrements of ill-health making it less prone to artificial ceiling and floor effects. The comparison of EQ-5D and SF-6D self-assessed health can be made using the information described previously in Table 2 and Table 3. The methods for valuing the health states are described later; what is important is that the instruments allow different ranges of scores as a result both of their construction and the method used for valuation. Each respondent completed both instruments before undertaking a further task. The health profile of each individual was then valued using pre-existing UK algorithms (Brazier, et al., 2002; Dolan, 1997), and plotted in Figure 10. Note that, as there are certain combinations of EQ-5D and SF-6D profiles that are shared by multiple respondents, each point in the the scatter plot is proportional in size to the number of people at that point (for example, the large data point at (1,1) identifies that there is a large group which answered that they were in full health for both instruments). To allow comparison, an additional black line is added to show points at which valuations under the two instruments are equal.

Figure 10: Self-Assessed Health Using the EQ-5D and the SF-6D



Before discussing these results, it is important to note that the divergence between the scores under the two instruments stems from both the construction of the instrument, and the method for valuing the health states combined within it. Konerding et al. (2009) investigated whether the former issue is significant, and concluded that the EQ-5D and SF-6D would “produce different valuations even if these valuations were determined according to the same principle” (p.1249). This point is echoed by Whitehurst and Bryan (2011), who argue that

“the descriptive classification systems differ to such an extent that contemporaneous EQ-5D and SF-6D valuations attached to health states should not be expected to provide similar estimates, irrespective of the preference elicitation technique used in the respective valuation studies.”(p.537)

There is clearly a high degree of agreement between the two instruments. Having stated there is some agreement between instruments, there are points of divergence which should be noted. Firstly, there is a large group of people for whom the EQ-5D does not identify ill-health (i.e. in Figure 10, those whose score under EQ5D_Dolan is 1) but the SF-6D does identify ill-health (in that their SF-6D_Brazier score is less than one). This group is shown by the red oval in Figure 10.

The second point of divergence is that there are a large number of observations with considerably lower EQ-5D scores than SF-6D scores. This reflects the floor effect which has been widely noted in the valuation of SF-6D health states (Brazier, et al., 2004). A floor effect can impact on the sensitivity of a MAUI to changes in quality of life for patients who are in very poor health, and inflate the cost per quality-adjusted life year (termed the incremental cost-effectiveness ratio, or ICER) for interventions in this type of patient group (Barton, et al., 2008; Stiggelbout, 2006).

Using Ordinary Least Squares (OLS) on the data presented in Figure 10, regressing the Brazier SF-6D score (y) on the Dolan EQ-5D (x) score gives $y = 0.445 + 0.407x$, with an adjusted R^2 of 0.577. However, the positive constant and coefficient on x being less than 1 indicate the relatively smaller spread of SF-6D scores, with predictable consequences for economic evaluation conclusions. These results are somewhat similar to those of a similar analysis by Khan and Richardson (2011) (whose OLS reports $y = 0.310 + 0.601x$), although they report a lower R^2 of 0.308. This difference might result from the relatively small sample size employed ($n=158$), or that the Khan and Richardson sample is based on a particular migrant group (Bangaldeshi) who were on average younger than our sample, and likely to differ both culturally and linguistically. Brazier *et al.* (2004) explore the relationship between self-assessed health using the EQ-5D and SF-6D. Using the UK preference algorithms, Brazier *et al.* regress SF-6D scores on EQ-5D scores for a large group ($N=2,436$) of patients across a range of conditions. In the simplest model, with the EQ-5D score being the sole independent variable, the coefficient on it was 0.33; this means that a change in health utility is expected to be measured to be three times larger in the EQ-5D than in the SF-6D. A similar pattern has been identified by others, including in people aged 45 and older (Barton, et al., 2008), and in mental health patients (Lamers, et al., 2006a). Thus, economic evaluations using EQ-5D to measure improvements in quality of life will, relative to those using the SF-6D, have larger incremental gains in quality of life, higher QALY gains, and hence lower ICERs. This is of increasingly greater importance in evaluations where the driver of the ICER is quality of life (rather than mortality), and suggests that the application of a common threshold for cost-effectiveness is problematic.

International valuation studies for the SF-6D have been conducted in a variety of countries, with publications based on populations from the United Kingdom (Brazier,

et al., 2002; Kharroubi, et al., 2007), Brazil (Gonçalves Campolina, et al., 2009), Japan (Brazier, et al., 2009), Portugal (Ferreira, et al., 2008) and China (Lam, et al., 2008). These used the Standard Gamble to value health states, with the inherent issues discussed in Section A of this chapter. Chapter 5 of this thesis will consider a novel approach to valuing the SF-6D.

Health Utilities Index (HUI)

The Health Utilities Index (HUI) was developed at McMaster University, and here refers to both the HUI Mark 2 and HUI Mark 3 classification systems (Horsman, et al., 2003). HUI1 was developed to evaluate outcomes for pre-term infants and is not generalisable to the broader community (Boyle, et al., 1983). HUI Mark 2 has seven dimensions, covering sensation, mobility, emotion, cognition, self-care, pain and fertility. It is notable that, relative to the EQ-5D and SF-6D, the dimensions are ‘within the skin’, meaning that the dimensions do not explicitly relate to how the respondent fits within society. Each dimension has between three and five levels and can describe 24,000 unique health states. HUI Mark 3 has eight dimensions (vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain) with five or six levels in each. The stated advantage of HUI3 over HUI2 lies in its additional descriptive ability, “*providing the potential for finer distinctions and enhanced responsiveness*” (Feeny, et al., 2002), (p.114). This seemingly small increase in dimensions and levels allows the HUI Mark 3 to describe 972,000 unique health states. The dimensions and levels are reproduced in Appendix 1. As there are a larger number of dimensions in HUI than in the EQ-5D and SF-6D, it is likely that the issue of structural independence is relevant here. Correlation between dimensions is highly likely; thus, the possibility for double-counting of aspects of ill-health in two or more dimensions is high. However, data comparable to that presented for the EQ-5D and SF-6D are not available.

An interesting aspect of HUI3 is that the worst level in the emotion dimension specifies a relativity to death, i.e. “(s)o unhappy that life is not worthwhile”. This relativity will become important in later parts of this chapter, as this description implies that any health state with this level ought to be valued as worse than immediate death, a condition they do not impose on the valuation of these health states.

While the instrument is likely to give good descriptive value, it might be argued that the extra descriptive ability is very rarely used. For example, the 1990 Ontario Health Survey asked 68,394 community-based respondents to complete HUI3; there were 1,755 unique health states represented in the data, meaning that less than 0.2% of the HUI3 health states actually occurred in the sample (Feeny, et al., 1995).

Assessment of Quality of Life (AQoL)

The Assessment of Quality of Life (AQoL) represents a program of work, which contains a variety of generic quality of life instruments. The original Assessment of Quality of Life (AQoL) instrument comprises 15 items, each with four levels, and it is this one which is replicated in the thesis appendix. All iterations of the instrument are available online at www.aqol.com.au. The 15 items in the original AQoL instrument are subdivided into five scales (illness, independent living, social relationships, physical senses and psychological well-being) (Hawthorne, et al., 1999). It was developed using a sample consisting of both hospital patients (n=143) and community respondents (n=111). Utility scores were generated using a larger sample (n=437) of the Victorian population with the Time Trade-Off (which will be considered in section B). It has undergone considerable validation, and population norms have been generated to provide a useful baseline for cost-utility analysis (Hawthorne and Osborne, 2005). Due to length, the AQoL instrument is not reproduced here, but is presented in Appendix 2 (Hawthorne, et al., 1999).

The AQoL team began with a definition of health, in which health is

“a state of optimum physical, mental and social well-being and not merely the absence of disease or infirmity.” (Hawthorne, et al., 1999) (p.210)

As part of a literature review, coupled with ongoing interviews of doctors iteratively feeding into the review, an item bank of potentially important components of health was developed. This item bank included both ‘within the skin’ issues and also ‘social experience’ issues. The AQoL team correctly argue that focusing on only one of these two groups is unlikely to capture the true health of the individual. A set of 138 general community members and one of 161 inpatients were asked to rate their current health state on each of the items in the item bank. Using a variety of instruments (including principle component analysis, exploratory factor analysis and structural equation modelling, an instrument was developed which the authors claim has some highly

attractive characteristics. These are that the dimensions are orthogonal, and each was uni-dimensional, and that there was high internal consistency.

Hawthorne *et al.* (2001) has compared it directly with other common quality of life instruments. They conclude that

“(i)t is concluded that at present no single MAU instrument can claim to be the ‘gold standard’, and that researchers should select an instrument sensitive to the health states they are investigating.” (p.358)

However, the comparisons they make are generally favourable to the AQoL in the dimensions they investigate including coverage and psychometric properties.

Regarding the most recent iteration of AQoL (the AQoL-8D), Richardson *et al.* (2011) identify that the AQoL has high test-retest consistency. While these conclusions are reasonable and reflect the significant efforts the AQoL team have made to ensure these characteristics hold in their instruments, the size of the instruments mean they contain a very large number of health states, cannot easily value more than a very small selection of them, and hence are constrained in their capability to investigate interactions and functional form.

One argument for using one of the smaller instruments is that, when administering an instrument to a patient population, the smaller instruments place less burden on the respondent. This is certainly true; for example, Richardson *et al.* (2011) note that the AQoL-8D takes almost 6 times longer to complete than the EQ-5D. However, this argument is undermined as the total time to complete even the most arduous instrument is negligible in the context of the wealth of data collection run in a trial setting. Richardson *et al.* (2011) state that the average completion time for the AQoL-8D is 5.5 minutes. I would argue that, in most settings, this is an acceptable burden if the investigator believes the quality of life scores from such an instrument are more reliable than those derived from a smaller instrument.

One final issue to consider when comparing and choosing the most appropriate instrument is whether the more important consideration is that the research community adopts one instrument as dominant, and uses this exclusively. This argument would result from asserting that what matters is consistency between economic evaluations, and has some merit. It eliminates the problem of gaming which

results from different instruments having different ranges of quality of life scores, and different sensitivity to changing quality of life in different dimensions (Kaplan, et al., 2011). Richardson *et al.* (2011) undertook a review of instrument use between 2005 and 2010, concluding that the EQ-5D was used in 63.2% of studies which employed a MAUI (relative to 14.4% for either HUI2 or HUI3, 8.8% for the SF-6D, and 4.3% for the AQoL). The argument that the EQ-5D should be universally adopted is plausible; however, it could be counter-argued that universal wrongness is not a sensible course of action.

Section B conclusion

This section has identified the major competing generic quality of life instruments. They differ in their construction, the number of dimensions and levels they contain, and the research penetration they have each achieved. Some common themes have emerged. The tension between breadth and depth of the instrument on the one hand, and the difficulty of valuing all health states has been identified with the latter being the focus of Section B. Structural independence has been identified as a major issue. If instruments violate this, it has implications for how health states are valued to avoid the issue of double-counting aspects of ill-health. Ceiling and floor effects have been considered in a comparison of the EQ-5D and the SF-6D, and reflected the insensitivity of the EQ-5D to relatively mild levels of ill-health.

Section C: Imputing values for other health states

Section A has described the dominant approaches for valuing individual health states. In all studies which attempted to value all health states within an instrument, a less than exhaustive set of states were valued in these ways; therefore, methods to impute other health states were required. Clearly, the importance of these methods increase with the number of health states that require this indirect valuation. The issue might be circumnavigated for smaller instruments which might allow direct valuation of all health states, but this has not been common practice to date. A recent Australian survey has attempted to produce an algorithm based on valuation of all plausible health states (Viney, et al., 2011b). However, this is not yet the dominant approach, and the method for imputing values remains an important topic (and indeed, the study by Viney *et al.* continued to adopt some of the imputation techniques to smooth the

results from their sample to ensure the monotonicity within dimensions that is built into the EQ-5D is maintained in the valuation of the health states).

Before critiquing the current approaches to imputing values for other health states, it is useful to identify that there is some developing research in the use of non-parametric Bayesian techniques to value health states, particularly in the context of the SF-6D. While the original research in the imputation of SF-6D health states focused on additive regression techniques, recent developments have suggested a non-parametric Bayesian approach might be more appropriate (Kharroubi, et al., 2005), and this approach was then employed using Brazier's existing data (Brazier, et al., 2002; Kharroubi, et al., 2007).

Kharroubi *et al's* (2007) approach claims a number of positive characteristics. These are the flexibility of the preference function, the previously mentioned monotonicity issue (which is less of a concern in the non-parametric approach), the value of perfect health being fixed at 1 without a considerable loss of predictive value and allowance for likely skewness (through the use of an exponential function).

Bearing this approach in mind, the attention of the chapter returns to the set of parametric approaches that have been taken to date. This focus on parametric approaches does not imply them to be superior to non-parametric techniques; rather the approach to analysis outlined in Chapters 3 and 4 is parametric and it is important to illustrate the characteristics of the current approaches which contrast with that taken in Chapters 3 and 4 and then in the subsequent empirical chapters.

Parametric approaches

Before considering the specifics of the parametric approaches to modelling taken by existing studies in the field, it is necessary to point out that the pre-specified functional form of the utility function is constrained by the choice of which health states within an instrument are directly valued by respondents. This issue is now discussed in the context of the EQ-5D, although the same criticism can be made of the standard approach to valuing SF-6D health states (Brazier, et al., 2002). For the EQ-5D, I have previously said that a recent study in Australia has considered all plausible health states (Viney, et al., 2011b); however, this is not generally the case.

Given that the EQ-5D has 243 individual possible states, it is unsurprising that no study has attempted to ask each respondent to directly value each of these states. Therefore, the pertinent question becomes how best to form a representative fraction of the entire space which allows a good estimation of the remainder of the EQ-5D states in whichever way that is defined. Prior to the Australian study of Viney *et al.* (2011b), two major approaches have been adopted to form this representative fraction. The original Dolan *et al.* approach (1997) valued 43 states, and each respondent directly valued a subset of these 43. An alternative approach (described here as the Tsuchiya approach) was developed which uses 17 states, all rated by each respondent (Tsuchiya, et al., 2002). The method by which these health states were selected is unclear. Dolan (1996) describes their approach,

“In choosing the states both for use in the study itself and for each respondent, the most important consideration was that they should be widely spread over the valuation space so as to include as many combinations of levels across the five dimensions as possible.” (p.143)

Tsuchiya argued that the 17 states directly valued in the Japanese valuation study were *“the minimum set of health states needed to estimate the value set”* (p.343). However, the minimum set of health states needed is dependent on the types of functional form that might be tested. If higher-order interactions are required, a much larger set would be required.

Lamers *et al.* (2006b) investigate these alternative approaches. Using data from Dolan *et al.* (1996), they assumed all respondents would value 11111 (full health) and in addition value 12,17,22,27,32,37 or 42 of the remaining 42 states. Samples of size 50, 100, 200, 300, 400, 600 and 800 were assumed. The outcome for each of these combinations is the mean absolute error (MAE) between the predicted values from the subsequent algorithm and the values observed in the data set. MAE is a useful tool for estimating appropriateness as it shows the fit of the model to the data. However, other diagnostics might also be of value, for example out-of-sample or split-sample prediction (of directly valued states or otherwise).

As expected, the MAE is negatively associated with both the sample size and the number of health states directly valued. Additionally, they contrast these data with the results of Dolan *et al.* (1995) which suggests that not only does the 17-state approach

used by Tsuchiya *et al.* (2002) lead to a lower MAE than that of Dolan *et al.* (1996) it may lead to a lower MAE than if each respondent valued 17 (or even 22) randomly assigned states from the 42 (although the difference does not appear to be statistically significant). The mean correlation for the predicted and actual values if 22 states from 42 are randomly selected is 0.986 (SD = 0.006) whereas the figures for the 17 states used by Tsuchiya *et al.* was 0.989 (SD = 0.002) (Tsuchiya, et al., 2002).

A related question concerns whether the 17 and 43 state approaches are optimal in terms of study design. To allow equal precision in each of the effect estimates, it is necessary to have equal frequency of appearance for each of the levels for each of the attributes. As there is a disproportionate number of the better health states, that is, states with attributes at level 1, in the 43 Dolan states (Dolan, et al., 1996) or the 17 Tsuchiya states (Tsuchiya, et al., 2002), there is greater precision at that healthy end of the scale. The other, related, issue involves the estimation of interactions. Although only 10 degrees of freedom are required for the estimation of main effects (i.e. one for each non-zero level of each dimension, a further 40 are required to estimate 2 factor interactions and of course if certain level combinations do not appear together (and perhaps do not make sense together) then estimation of all two factor interactions becomes impossible.

Additivity (e.g. EQ-5D and SF-6D)

The first major issue to address is how best to specify a functional form for the imputation. It is necessary to define a utility function over which the observed choices are made, confirm this utility function performs relatively well in terms of reflecting observed valuations of health states, and then extend the utility function to estimate quality of life scores for all health states defined by the quality of life instrument.

The original research in both the EQ-5D and the SF-6D focused on a predominantly additive utility function (Brazier, et al., 2002; Dolan, 1997); therefore the disutility of a dimension moving to a worse level is assumed to be constant irrespective of the levels of other dimensions.

This might be related to the previously discussed idea of structural independence, in that the latter is likely to be a necessary but not sufficient condition for an additive utility function. If two dimensions are measuring the same disutility, then the impact of having both of them at a poor level is reasonably likely to be less than the sum of

the impact of either being at a poor level (thus a positive coefficient on the interaction term). However, structural independence does not ensure a purely additive utility function as conventional interaction terms can still apply.

The Measurement and Valuation of Health (MVH) group at the University of York were the first to produce a full valuation set for the EQ-5D, published by Dolan (1997). In the preferred algorithm, Dolan's consideration of interaction terms is limited to an N3 dummy term. This is a dummy variable equal to one if any of the five EQ-5D attributes is at the worst level. Thus, the score Y an individual gives to a health state is defined as:

$$Y = \alpha + \beta_1 MO + \beta_2 SC + \beta_3 UA + \beta_4 PD + \beta_5 AD + \beta_6 M2 + \beta_7 S2 + \beta_8 U2 + \beta_9 P2 + \beta_{10} A2 + \beta_{11} N3 \quad \text{Equation 10}$$

Where MO, SC, UA, PD and AD are equal to 1 if the dimension is at level 2, and equal to 2 if the dimension is at level 3, and M2, S2, U2, P2, A2 are equal to 1 if the dimension is at level 3. This can easily be re-specified to provide more intuitive coefficients by dummy coding MO, SC, UA, PD and AD to be equal to 1 if the dimension is at level 2, and M2, S2, U2, P2, and A2 equal to 1 if the dimension is at level 3.

The interpretation of the N3 term is difficult. If a dimension (for example, mobility) moves to level 3 from level 1 (through an active person becoming confined to bed), the decrement to utility is $2\beta_1 + \beta_6 + \beta_{11}$. If another dimension is already at level 3, the decrement to utility of becoming bed-ridden is $2\beta_1 + \beta_6$. Thus, any second or subsequent dimension to move to level 3 has a smaller decrement in utility than the first (as β_{11} is positive in every existing algorithm).

While Australian data suggest that the main area for interaction effects is between dimensions at level 3 (Viney, et al., 2011b), Dolan's approach is somewhat blunt in that it constrains the interaction term to be constant across any pair of level 3 dimensions. However, as discussed previously, the data they generated were limited in terms of the types of pairwise interactions they were able to estimate.

It should be noted that Dolan considered a variety of alternative specifications involving interactions. These interactions were the N3 term previously described, the products of each main effect to allow investigation of first-order interactions, and

dummies for when 1,2,3 or 4 of the dimensions were at level 1, or when 1,2,3,4 or 5 of the dimensions were at level 3. None of these significantly model fit other than the N3 term, however it can be argued this is a consequence of the states directly valued (Viney, et al., 2011b). One issue with the interactions Dolan considered was that the dummies designed to capture first-order interactions were unusual. Because of his coding described above, the first-order interaction was between any pair of MO, SC, UA, PD and AD (meaning ten possible interactions). However, this imposes that the interaction effect when (for example) SC = 2 and UA = 1 (i.e. at levels 3 and 2 respectively) is the same as that when SC = 1 and UA = 2 (in which the order is swapped). This is by no means an obvious conclusion, and impedes the consideration of interactions.

For the SF-6D, the original approach taken by Brazier *et al.* (2002) used a similar additive approach. His preferred algorithm is modelled on the mean scores of the 249 health states directly valued by respondents (so an observation is the mean score for a particular directly-valued state rather than one valuation made by a respondent). It has a dummy for each level other than the best in each dimension, plus a MOST dummy variable which is equal to 1 if and only if a dimension is at the worst level (which is analogous to Dolan's N3 dummy term).

Unlike Dolan's EQ-5D algorithm, Brazier *et al.* do not allow a freely estimated constant in their preferred model. Thus, the regression forces full health to be valued at 1 in the regression, rather than imposing it subsequently (as was done by Dolan). This is an important distinction in milder health states. Contrasting Brazier's preferred algorithm with the most similar model with a freely estimated constant (model 8 in their paper), the constant is 0.788. The LEAST dummy variable (which works in the same way as MOST, but with dimensions at the best level) is also 'turned on' and valued at 0.048, so the baseline from which disutility of health states is $0.788 + 0.048 = 0.836$. This is important because moving from the mildest health state to full health in the unconstrained model 8 improves quality of life by 0.148, while the comparable figure for the constrained model is 0.011.

Multiplicativity (e.g. Health Utilities Index (HUI))

Both HUI and AQoL employ multiplicative methods for extrapolating values to health states within their instruments which are not directly valued by respondents. The

approaches are similar, hence the methods for one (HUI) are outlined now; for details on the methods employed to value the original AQoL instrument, see Hawthorne *et al.* (Hawthorne, et al., 2000). The HUI scoring system is based on a multiplicative approach (Feeny, et al., 2002), and is derived using the Standard Gamble. In this approach, the utility u of health state x is estimated by

$$u(x) = \frac{1}{k} \left[\prod_{j=1}^n (1 + k_j u_j(x_j)) - 1 \right] \quad \text{Equation 11}$$

where

$$(1 + k) = \prod_{j=1}^n (1 + k_j) \dots \quad \text{Equation 12}$$

in which $u_j(x_j)$ is the single attribute utility function for attribute j , and k and k_j are model parameters. The $u_j(x_j)$ term places each level of each dimension on a scale with the worst level at zero and the best level at one. The k_j term is the relative importance of each dimension (and hence the importance of the single attribute utility function scores for that dimension in valuing a multi-attribute health state). The k captures the interaction in preferences among attributes. If k is positive, attributes are preference complements. Conversely, if it is negative, attributes are preference substitutes. This approach is described by Keeney and Raiffa (1993).

Before presenting the results, it is important to note that the wording of HUI3 produces an oddity in the valuation algorithm. In the Emotion dimension, Level 5 states that the respondent is “*so unhappy that life is not worthwhile*”. It is arguable therefore that the valuation of any health state with this level should be less than zero; however, this is not applied in the valuation study.

The single attribute utility function scores are not presented here; however, they can easily be inferred from the information below. Table 8 provides the coefficients associated with the multi-attribute utility function.

Table 8: HUI3 Multi-Attribute Utility Function

Level	Hearing	Speech	Ambulation	Dexterity	Emotion	Cognition	Pain
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	0.98	0.94	0.93	0.95	0.95	0.92	0.96
3	0.89	0.89	0.86	0.88	0.85	0.95	0.90
4	0.84	0.81	0.73	0.76	0.64	0.83	0.77
5	0.75	0.68	0.65	0.65	0.46	0.60	0.55
6	0.61		0.58	0.56		0.42	

Source: Feeny 2002 (2002)

where $u = 1.371$ (product of coefficients on levels) – 0.371. The method for estimating the health score attached to a particular state therefore assumes that the negative impact of a poor level of a particular dimension depends on the levels of other dimensions. The multiplicative approach suggests the impact of worsening health is highest when other dimensions are at relatively better levels. Thus, the maximum value is 1, and the minimum is -0.351. As an example of the impact of interactions, a movement from level 1 hearing to level 6 hearing can be considered. If the other six dimensions are at level 1, the impact on quality of life due to this reduced hearing is to reduce quality of life by 0.53 (from 1.00 to 0.47). If the other six non-hearing dimensions are at level 4, the impact is 0.10 (from -0.12 to -0.22). If everything other than hearing is at the worst level, and then hearing falls from level 1 to level 6, the impact is smaller than in the previous cases, with quality of life falling by 0.01 (from -0.34 to -0.35). This structure imposes a quite specific restriction on utility functions. While the existence of interactions of this sort are plausible, the multiplicative structure imposes the notion that the impact of a deteriorating dimension depends on the total disutility of all other aspects of quality of life, rather than on individual aspects of that pre-existing disutility.

Algorithms have also been generated for HUI2 in the United Kingdom (McCabe, et al., 2005), and for HUI3 in France (Le Galés, et al., 2002). The results of these two studies differed relative to the Canadian algorithms. The French results were similar, the UK ones were not.

Chapter conclusion

In this chapter, I have considered three areas. The first section considered how health states are valued. The methods used for valuing specific health states were discussed. This included discussion of Time Trade-Off, Standard Gamble and Visual Analogue

Scales. The former two were preferred to the third as they are based on a notion of sacrifice. However, problems with each were identified relating to cognitive challenge of answering questions, and the effect of respondent characteristics beyond health preferences, such as attitude to risk and time preference, that impact on the valuation of health states.

The second section in this chapter considered how health is described in generic quality of life instruments. The concepts of breadth and depth were introduced, both of which add to the sensitivity of the instrument. However, this extra sensitivity comes at the cost of increased difficulty in valuing all health states within the instrument accurately. This trade-off means that the choice of quality of life instrument is context-specific, depending on the expected health changes in the population under consideration.

The third section looked at the methods employed for imputing scores for non-directly valued health states. This is of particular concern for those instruments in which a relatively small proportion of health states were directly valued by respondents. The imputation methods to date are generally sound; however, they are often restricted by the selection of directly valued health states. The choice of health states has a direct impact on the parameters (and interactions between parameters) that can be investigated, and this is something lacking from the majority of the existing literature in the area.

An alternative approach which will be introduced in the next chapter is the use of ordinal data in the form of the discrete choice experiment. This technique will be tested for the SF-6D in Chapter 5. The issues concerning specification of functional form are similar to those presented here. As with the additive and multiplicative structure described here, it will prove important to choose a preferred functional form prior to designing the fieldwork. While the issue is common to the task presented in this chapter, and the next, the advantage in this area of the DCE is that the techniques used to produce an appropriate design of the experiment are probably further advanced and more widely considered.

Chapter 3: Discrete Choice Experiments: Principles and Application for Health Gain

Chapter summary

This chapter introduces the concept of the discrete choice experiment (DCE), and discusses the strengths and ongoing controversies associated with this approach. It is a stated preference (SP) approach, the increasing use of which reflects the difficulty in collecting revealed preference (RP) data in most health contexts. The technique provides an alternative to other SP approaches such as the Time Trade-Off, Standard Gamble and Visual Analogue Scale, offering attractive solutions to some of the problems identified in Chapter 2. Chapters 5 and 6 will use the DCE to explore some of the issues raised in Chapters 2 and 1 respectively. In the Standard Gamble, Time Trade-Off and Visual Analogue Scales, the respondent is asked to quantify their strength of preference regarding each health state being valued. In DCEs, the survey response is to simply identify which of a set of options is preferred (thus, it is ranking-based). In this chapter, existing ranking-based approaches in the health valuation field are critiqued. It is argued that they are flawed as the weights they produce do not capture the trade-offs between time (and / or probability) and the other components of the choice experiment (for example the quality of life terms required for the QALY model). If we are interested in these trade-offs, it follows that an appropriate numeraire (e.g. time, probability) must be part of the choice set faced by the survey respondent. Following this, I will discuss the existing gold-standard methods for analysing choice experiment data. This includes the consideration of respondent heterogeneity (both on observed and unobserved characteristics), and some thoughts about how this heterogeneity might be considered within health policy decisions. Following this, I discuss a method for adapting the DCE approach to account for the valuation of health gain due to the complementarity of time and other variables such as quality of life.

The existing methods for estimating welfare measures from choice experiments are then discussed. The two leading contenders for this, compensating variation and marginal rates of substitution, are outlined, along with the types of situations in which each might be more appropriate. An approach close to that used when estimating

marginal rates of substitution is then proposed for use in Chapters 5 and 6, which is defined by the ratio of marginal utilities.

Introduction

In the preceding two chapters, I have introduced the concept and necessity of economic evaluation in healthcare, and investigated how quality of life is most usually described and valued for the purpose of constructing an outcome measure combining both mortality and morbidity. I have attempted to illustrate some of the constraints and assumptions that are made concerning how society ought to value different types of health gain. The surveying of individuals within a society can help make this process more representative of societal preferences, and hence relax some of these assumptions. One area in which this surveying can play a role is in this description and valuation of health. Another area in which it can be useful is in testing whether the assumption of inequality neutrality, which is implicit in the common approaches to economic evaluation of healthcare, holds. These are the two areas addressed by the empirical chapters in this thesis. In this chapter, I will describe the method used in this thesis for investigating both of these issues. This is the discrete choice experiment (DCE), a type of stated preference experiment with several attractive characteristics. However, I will argue that care has to be taken in applications of this methodology in this setting, as DCEs have to be adapted to reflect the unusual co-dependence of quality of life and life expectancy.

Stated and revealed preference data

For economists, data for the modelling of behaviour are generally derived from real choices; the choices observed in these data are likely to be good predictors of future behaviour, and are less likely to be subject to possible problems such as interview bias. In the health sector, areas exist where such data cannot be collected. For example, we cannot readily use revealed preference data to investigate how much people would be willing to pay for a health resource if we are in a healthcare system in which people are never faced with a non-zero price. Economists are interested not only in areas where revealed preference data are plausible, but also in goods and services which either are not traded, or have not yet been traded. Therefore, revealed preference data are of use only in certain domains of health, none of which are addressed in this thesis.

The concept of stated preference as a tool to elicit social values has a long history, dating back at least as far as Thurstone's attempt to identify the relative importance weight of a selection of contemporary crimes (Thurstone, 1927a). Discrete choice experiments are an attribute-based approach to collecting stated preference data. It is based on the Lancasterian approach, in which the preference associated with a choice is a function of various attributes of the option, and of nothing else (Lancaster, 1966). When faced with a choice between multiple options, it is assumed that the respondent will select the option which maximises their utility (this was discussed previously in the context of welfarism and will be formalised in this chapter).

The principle behind the methods discussed in this chapter begins by stating that a set of factors x can explain a choice Y within a stated preference framework,

$$P(Y = 1 | x) = F(x, \beta), \quad \text{Equation 13}$$

and, since the only values Y can take are 0 (not chosen) and 1 (chosen), it follows that

$$P(Y = 0 | x) = 1 - F(x, \beta) \quad \text{Equation 14}$$

The coefficients β reflect the impact of changing levels of x on probability of choice. The question is how to specify F ; an obvious starting point is to assume a linear probability model,

$$F(x, \beta) = x' \beta \quad \text{Equation 15}$$

However, as noted in Greene (2003), this does not assure the analyst that model predictions are sensible as probabilities (i.e. they can fall outside of a 0-1 range). The commonly applied solution is to adopt a continuous probability distribution, such as the standard normal or logistic distribution. Under a normal distribution, a probit model is derived in which

$$P(Y = 1 | x) = \int_{-\infty}^{x'\beta} \phi(t) dt = \Phi(x'\beta) \quad \text{Equation 16}$$

If a logistic distribution is assumed, a logit model is derived in which the corresponding probability distribution is

$$P(Y = 1 | x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}} = \Lambda(x'\beta) \quad \text{Equation 17}$$

where $\Lambda(\cdot)$ is the logistic cumulative distribution function.

Rather than modelling probability, it is standard for discrete dependent-variable models to be considered in terms of index function models. Thus, what is observed is the choice of the respondent, and we assume this to be driven by some latent utility function where the option which is most likely to be picked by the respondent is that with the highest utility. The reason why the option with the highest utility has the highest probability of being picked, rather than the certainty of being picked as it maximises utility is that the latent utility function includes an error term ε with mean 0 and a variance of either $\pi^2/3$ in the logit case or 1 in the probit case. Thus, individual choice behaviour is assumed to be intrinsically probabilistic (Marschak, 1960; McFadden, 1974).

The utility of an alternative i in a choice set C_n to an individual n is given by

$$U_{in} = V(X_{in}, \beta) + \varepsilon_{in} \quad \text{Equation 18}$$

The $V(X_{in}, \beta)$ term is the explainable (or systematic) component of utility which is determined by characteristics of the choice or the individual n . However, there is also an error term which differs over alternatives and individuals and makes prediction of choice uncertain. It is assumed that the individual will choose the option if the utility associated with that option is higher than any alternative option. If we assumed there are J items in C_n , the choice is defined as

$$y_{in} = f(U_{in}) = 1 \text{ iff } U_{in} = \max_j \{U_{ij}\} \cdot \forall j \neq i \in C_n \quad \text{Equation 19}$$

Alternative i is chosen if and only if

$$(V_{in} + \varepsilon_{in}) > (V_{jn} + \varepsilon_{jn}) \cdot \forall j \neq i \in C_n, \quad \text{Equation 20}$$

which can be rearranged to yield

$$(V_{in} - V_{jn}) > (\varepsilon_{jn} - \varepsilon_{in}) \cdot \forall j \neq i \in C_n \quad \text{Equation 21}$$

The role of Random Utility Theory

Neither the systematic nor the error components in the utility function are directly observed. Therefore, analysis is reliant on observing choices and inferring the terms from that. Random Utility Theory (RUT) is the dominant approach to doing this. In RUT, it is assumed that the difference in utility between two options (in this case i and

j) is proportional to the frequency that one is chosen over the other (McFadden, 1974; Thurstone, 1927b).

For this to hold, it is necessary to assume that the variance of the random error term is constant across choices and individuals. This is because, if an individual has a weak preference for i over j , but answers very consistently (i.e. has a low variability on the random error term), assuming a higher variability for that person makes i appear strongly preferred to j . Assuming that variance is known is described by Greene as an 'innocent normalisation' (p.669)(2003); this is generally true but has to be considered when discrete choice experiment data are analysed (which will be discussed later in the context of the Scale Multinomial Logit model).

Before looking at these issues, two areas of importance regarding the use of discrete choice experiments to value health gains need to be outlined, and will impact on how I decide to specify the latent utility function in the subsequent empirical chapters. These issues are the inclusion of an appropriate numeraire in the discrete choice experiment, and the inclusion of variables in the analysis not as main effect terms, but interacted with the numeraire. These two issues will now be addressed.

A suitable numeraire

In the QALY model, the value of a chronic health profile (i.e. one with constant quality of life until death) was described in Figure 2. Assuming we are considering the health profile of one individual, the value of a health profile is simply the product of length of life and quality of life. A more general framework which explains both the Time Trade-Off and the Standard Gamble would be to state that, under Expected Utility Theory, the value of an uncertain health profile is the product of length of life, quality of life and probability of receiving that profile:

$$\text{Value} = \text{Life_Expectancy} * \text{Quality} * \text{Probability} \quad \text{Equation 22}$$

The value of a gamble involving multiple possible health states is then simply the sum of the values of the possible health states (which are automatically weighted by the likelihood of being received).

Under both the TTO and Standard Gamble approaches described in Chapter 2, this function is used, and assumes that people are risk-neutral and have a linear utility function with respect to time (although each technique only explicitly considers one

of these two constraints). In the Time Trade-Off, a value for quality of life for a particular health state was determined by keeping probability constant (at 1), and finding a point at which the individual is indifferent between ten years in the health state being valued, and some shorter period in full health. Similarly, in the Standard Gamble, a value for quality of life for a particular health state was determined by keeping life expectancy constant (although this is not always explicit), and then finding a point at which the individual is indifferent between a 100% likelihood of the health state being valued, and a less than 100% chance of full health (and a complementary probability of death). The important point to make is that both the Standard Gamble and Time Trade-Off include a concept of sacrifice; both involve trading (be it risk or time) for a better quality of life.

In the existing literature attempting to use RUT to estimate utility weights, this issue of a suitable numeraire has been ignored. The recent examples of studies which have attempted to use ordinal data and Random Utility Theory to produce QALY weights have assumed that it is possible to extrapolate from a simple ranking of health states and death to a set of weights (McCabe, et al., 2006; Salomon, 2003). Stolk *et al.* (2010) have similarly produced utility weights for a multi-attribute utility instrument (the EQ-5D, which was discussed in Chapter 2), but using a DCE rather than conventional rank data, but again ask respondents to trade between profiles consisting of health profiles independent of duration. Flynn *et al.* (2008) argue that the methods employed by these studies have been erroneous for the following reason.

In the approach taken in these three studies, neither length of life nor probability is a factor in the choice; thus the rankings in McCabe *et al.*, or Salomon, and the DCE of Stolk *et al.* do not include a numeraire required for the QALY model. Within these pre-existing studies, it is not possible to impose this multiplicative structure between quality of life and life expectancy (and probability) on the scores for particular health states. While the results from these studies seem plausible, a value on a health state of 0.5 (i.e. halfway between full health and death), does not necessarily imply that a person is willing to sacrifice half of their existing life to move from this health state to full health, or to accept a 50% risk of death to return to full health. Thus, it is essential that a suitable numeraire be included in the choice experiment; in this thesis, time is assumed to be this suitable numeraire. However, it is plausible that risk might be used

as an alternative, in a DCE analogous to the Standard Gamble technique introduced in Chapter 2.

Choice experiments and health gain

A common approach to specifying the systematic component of the utility function in the analysis of choice data is to include a main effect for each level of each attribute (with one level in each attribute omitted to avoid over-identification). This approach to the utility function can be extended to consider interactions between levels of attributes. In the case of health gain, considering quantity of life separately from characteristics of those extra years (such as quality of life), and including simple interactions between them is unlikely to be adequate in capturing this inter-relatedness. The reason for this is because it is necessary to impose the zero-condition discussed in the introductory chapter in the context of QALY construction. Specifically, the utility of a health profile with zero life expectancy is zero irrespective of the quality of life in that (non-) period (Bleichrodt, et al., 1997).

If we assume that the utility function for all models is linear with respect to time (denoted by the *TIME* variable), and the characteristics of that time enters not as a main effect, but as an interaction with the *TIME* variable. This is denoted as **Utility Function 1**, which will be used for Models A, B and C in Chapter 5 (and Model A in Chapter 6). Thus, the utility of alternative *j* in scenario *s* for individual *i* is

$$U_{isj} = \alpha TIME_{isj} + \beta X'_{isj} TIME_{isj} + \varepsilon_{isj} \quad \text{Equation 23}$$

The marginal utility of time in this approach, which will be used later, is therefore

$$\frac{\delta U_{isj}}{\delta TIME_{isj}} = \alpha + \beta X'_{isj} \quad \text{Equation 24}$$

A major advantage of this approach is that, if *TIME* is set at zero, the systematic component of the utility function is zero. Thus, the choice between two profiles with *TIME* set to zero is random irrespective of the levels of the other parameters. In terms of the constraints the QALY model places on individual preferences, it is noteworthy that Equation (23) features both conditions specified by Bleichrodt, Wakker and Johannesson (1997) as being jointly sufficient for the QALY model discussed in the introductory chapter. Utility is linear with respect to time (which is imposed by risk neutrality), and the utility function passes through the origin by construction. Thus,

this is the QALY model. As discussed in Chapter 1, the zero-condition is uncontentious. As noted in Chapter 2, the appropriateness of imposing linearity of utility with respect to time is considerably more uncertain; is ten years in full health associated with double the utility of five years in the same state? As the more contentious of the two requirements for the QALY model, the linearity of utility with respect to time will be relaxed (with that relaxation being tested) in both empirical chapters, as described below.

Utility Function 2, which is used for Model D in Chapter 5, (and Model B in 6) builds on Utility Function 1 by relaxing the restriction of linearity of utility with respect to time. Thus, Utility Function 2 is:

$$U_{isj} = \alpha TIME_{isj} + \rho TIME_{isj}^2 + \beta X'_{isj} TIME_{isj} + \phi X'_{isj} TIME_{isj}^2 + \varepsilon_{isj} \text{ Equation 25}$$

The corresponding marginal utility of time is therefore

$$\frac{\delta U_{isj}}{\delta TIME_{isj}} = \alpha + \beta X'_{isj} + 2\rho TIME_{isj} + 2\phi X'_{isj} TIME_{isj}, \text{ Equation 26}$$

which can be simplified to

$$\frac{\delta U_{isj}}{\delta TIME_{isj}} = \alpha + \beta X'_{isj} + 2TIME_{isj}(\rho + \phi X'_{isj}) \text{ Equation 27}$$

Thus, the linearity of utility with respect to time is relaxed, as reflected in the $\rho TIME_{isj}^2$ term in Equation (25). However, the inter-relatedness of duration and the other terms remains. Additionally, it relaxes the assumption that the disutility associated with the X'_{isj} terms is independent of time (the $\phi X'_{isj} TIME_{isj}^2$ term). Using a QALY example, if a high level of pain is worse than a high level of anxiety over 5 years, the same relativity is necessarily true over any other number of years within the QALY model. This is imposed, rather than reflective of any data, and should be tested. With regard to the marginal utility of time, an important point to note is that it depends on the level of *TIME*; this has implications for the estimation of welfare measures, which will be undertaken in the two empirical chapters. Specifically, as Equations (25-27) include a *TIME* term, it means that the analysis does not impose the constraint that marginal utility of time is constant across values of *TIME*.

Up to this point, the chapter has introduced discrete choice experiments as a viable method for collecting and analysing stated preference data. The next step is to address a number of additional issues which shape the structure of the empirical chapters to follow. These are the methods for dealing with lexicographic preferences (which violate random utility theory (RUT)), and the importance of respondent heterogeneity. A further issue, that of the appropriate design strategies to ensure mathematically efficient data collection and unbiased results, will be described in depth in Chapter 4.

Lexicographic preferences surrounding death

Flynn (2010) claims that it is problematic to consider death within a random utility theory framework. His argument is that there is a proportion of respondents who will never acknowledge a health state to be less preferred than immediate death. Of this group, some may just not see a health state they believe to be worse than immediate death. Others however, might believe that it is

“not for humans to decide that death is preferable to a living state, no matter how bad it is.” (p.3)

If this is the case, Flynn argues that these people violate RUT, under which there is always a non-zero probability of an individual picking an option in a choice set (and these people will never select death). This violation means they must be excluded from the dataset. Is this a valid critique? It is certainly true that there may be people with these preferences, and importantly and more troublesome from an analysis perspective, that it is difficult to identify whether someone who never selects death is in this group or not. One possible counter-argument is that this type of lexicographic preference is likely to still exist when people respond to a choice set in which there is no death option. In the context of the EQ-5D discussed in the previous chapter, it is possible that someone might never pick an option which involves being confined to bed (which is the worst level in the Mobility dimension). It is uncertain if Flynn’s position extends to excluding these people from further analysis; however, it is logically difficult to assert this is a different type of lexicographic preference, only that it may be less likely than a refusal to prefer death over some non-death profile. Nevertheless, the analysis undertaken in the empirical chapters of this thesis are not reliant on preferences relative to an immediate death option (although the empirical

work presented in Chapter 5 was built to allow analysis using rankings between health profiles and death).

Modelling respondent heterogeneity

The next stage of this chapter is to consider approaches to the modelling of heterogeneity. This may seem tangential to the estimation of QALY weights as QALY weights are population average weights. However, there are two reasons why heterogeneity modelling remains potentially interesting in this instance. Firstly, as noted by Hensher *et al.* (1999), and by Swait and Louviere (1993), aggregation of estimates from discrete choice tasks can take place only after variance heterogeneity is accounted for. This does not occur in the base case analysis proposed below (or in a conditional logit or a mixed logit, also described below). The second reason is that it is of some interest to know the degree of harmony with which society holds the mean view. If there is significant disagreement about the value of health gain, and the elements of health gain that are of most importance, the use of a 'one size fits all' model may be considered less appropriate.

The modelling of heterogeneity is divided into two components. Firstly, it is possible to identify heterogeneous results based on observable characteristics of respondents. Conversely, it might be that modelling of response heterogeneity not based on observable characteristics may be valuable. The first is discussed now, and the latter is described at length subsequently.

Observable characteristics and heterogeneous responses

Conventional valuation of generic health states for use in economic evaluation is focused on the mean respondent. This is reflected in the valuation of health states within the MAUIs discussed in the previous chapter. This is appropriate within the convention that the value attributed to a health state is a societal one. However, it is useful to investigate whether respondents completing this type of survey differ in their responses based on observable characteristics. This is useful for two reasons. Firstly, it will identify the importance of using a balanced (i.e. population-representative) panel; if respondents do not differ in predictable ways, it is relatively less important that a balanced panel is used. The second reason is that it is intrinsically interesting to explore the degree to which people agree with the mean respondent.

Considering each demographic characteristic in turn, the sample can be split into mutually exclusive and exhaustive groups; for example, those above and below median income, males / females etc. The utility function of alternative j for individual i with or without demographic characteristic c in scenario s is

$$U_{icsj} = X'_{isj}(\beta_i + \beta_c) + v_i + \varepsilon_{isj} \quad \text{Equation 28}$$

The v_i term is an individual-specific error term and will be discussed more later in the context of the base case random-effects probit model. With regard to the sub-group analysis, each c was run separately, and this was then compared with a pooled model, and the importance of the c terms were investigated using Information Criteria, and by testing for poolability using a likelihood-ratio test. The results are presented graphically by rescaling the results from each sub-group such that the coefficient on duration is set to 1 (thus placing the results from the subgroups on a common scale). If, once the adjusted values are generated for two mutually exhaustive sub-groups, there is a difference between the two for a particular level of a particular dimension, it means that the amount of life expectancy that an individual is willing to sacrifice to move to full health in that dimension differs.

Modelling heterogeneity on unobservable characteristics

Having outlined the general method adopted for considering the modelling of heterogeneity based on observable characteristics, the chapter now turns to modelling of unobserved heterogeneity. The approach taken is adapted from that used by Fiebig *et al.* (2010). Basically, the analysis approach described here and employed in Chapters 5 and 6 uses a random-effects model for the base case, and then consider a range of modelling approaches which incrementally build on each other towards the very flexible generalised multinomial logit model. The description of the approaches begins with the most constrained approach taken, then relaxes assumptions one at a time. Each of the more constrained models are nested in one of the subsequent models. The labelling of the heterogeneity exploration models reflects the number of parameters estimated under each model (A being the least and F being the most). The base case random-effects model is inserted into this as the second most constrained approach.

Note that the general approach in both empirical chapters is to pair the utility function specified in Equations 23 and 25 (i.e. with and without the imposition of linearity of utility with respect to time) with each of the 7 models specified below i.e. the random-effects probit, the conditional logit, the scale multinomial logit, the mixed logit (with and without correlation) and the generalised multinomial logit (with and without correlation). Each of these will now be described, in order of the number of degrees of freedom starting from the most constrained.

Conditional logit modelling (heterogeneity exploration model 1)

I have previously noted that, when there are only two choices for the respondent, binary choice models can be employed. The reason for using these approaches in place of standard Ordinary-Least Squares (or similar) linear probabilities models is that, while the latter are possible, they are somewhat limited in that they cannot guarantee predicted probabilities in the appropriate range, and in that the error term is heteroskedastic, depending on the β values (Greene, 2003). The methods described below improve on this linear probability modelling by ensuring that the probability of an event (in this case an option being chosen) lies between 0 and 1, and also that the partial effect of an explanatory variable can differ based on other explanatory variables (Wooldridge, 2003). The latter is important because changing the level of a dimension within an option in a choice set is likely to have a very different impact on the probability of selecting that option if the base probability is 0.5 or 0.95. The binary choice models are based (as are more advanced specifications such as mixed logit, and generalised multinomial logit which will be described later) on the random utility model described by McFadden (1981). The latent utility function for individual i of alternative j in scenario s is defined in the following way:

$$U_{isj} = X'_{isj} \beta_i + \varepsilon_{isj}, \quad \text{Equation 29}$$

where β_i is a vector of co-efficients and X_{isj} is a vector of explanatory variables. If we assume the error term to be identically and independently distributed (iid) as extreme value, we generate the standard general multinomial logit (MNL) specification in which the probability that the individual chooses alternative j in scenario s is defined as

$$P_{isj} = \frac{\exp(X'_{isj}\beta_i)}{\sum_h \exp(X'_{ish}\beta_i)} \quad \text{Equation 30}$$

As there are only two choices in the binary choice model, we can define k as the alternative which is not j and then simplify this to

$$P_{isj} = \frac{\exp(X'_{isj}\beta_i)}{\exp(X'_{isj}\beta_i) + \exp(X'_{isk}\beta_i)} \quad \text{Equation 31}$$

The conditional logit has a number of constraints which may be unrealistic. When predicting responses, it forces the data to conform to Independence of Irrelevant Alternatives (IIA). This implies proportional substitution patterns (i.e. if an option drops out of a choice set, the respondents who initially selected that option are reallocated to the other options proportionally to the probability of these options being selected in the initial choice set). For the binary choices considered in this thesis, this does not pose a problem.

Base case analysis - Random-effects (RE) modelling

The conditional logit outlined above assumes all observations are independent; hence, there it has no capability for reflecting the panel nature of data. Responses from a common individual are likely to have a degree of commonality. To assume independence of each response is to exaggerate the degree of agreement in the sample. In the random-effects (RE) probit, the error term is a composite term, consisting of a standard error term ε_{isj} distributed iid standard normal, and a person-specific error term v_i (distributed iid normal with mean 0 and variance σ_v^2)⁸. Thus, Equations (23) and (25) are amended by stating the utility of alternative j in scenario s for individual i is

$$U_{isj} = \alpha TIME + \beta X'_{isj} TIME + v_i + \varepsilon_{isj}, \quad \text{Equation 32}$$

and the less constrained utility function allowing for non-linearity over time becomes

$$U_{isj} = \alpha TIME + \rho TIME^2 + \beta X'_{isj} TIME + \phi X'_{isj} TIME^2 + v_i + \varepsilon_{isj} \quad \text{Equation 33}$$

⁸ The move between logit and probit functions is defined by the characterisation of the random error term. In a logit function, it is distributed following a standard logistic distribution. In a probit model, it is assumed to be standard normal. The justification for moving to a probit model for the random-effects approach is that random-effects logit models are rarely used in practice

Thus, as the person-specific error term applies across all choices made by an individual, these choices will not be independent. The correlation between choices made by an individual, which is a check of whether there are person-specific effects, is estimated as

$$\rho = \sigma_v^2 / (1 + \sigma_v^2) \quad \text{Equation 34}$$

In the empirical chapters to follow, analysis will follow two directions. The base case analysis will be undertaken using the random-effects probit (the *xtprobit* command in STATA)⁹. It accounts for the panel nature of the data produced by the discrete choice experiment methodology, but does so in a fairly parsimonious way. The reason for using probit rather than logit is simply reflective of the more common usage of it in the economics literature; in all empirical chapters, the base case model will also be run using *xtlogit* to investigate whether the results are sensitive to this switch (and thus able to be compared with the various heterogeneity exploration models which use a logit specification).

Heterogeneity exploration model 2 - Scale Multinomial Logit modelling

In the conditional logit, the error term has a variance that has been normalised to one in order to achieve identification (Fiebig, et al., 2010). This is often termed as a perfect confounding of the estimates of the mean and variance of the latent utility scale (Ben-Akiva and Lerman, 1985; Swait and Louviere, 1993). If a respondent is relatively consistent in her choices, her individual error term is relatively small. Since the variance of the error term is normalised, the individual-level coefficients for the respondent will tend to be universally higher because of this consistency. To introduce (and control) this scale term, the utility function for individual *i* of option *j* in choice set *s* is therefore

⁹ Under the *clogit*, *mixlogit*, and *gmnl* commands that form the strands of the heterogeneity exploration component of the empirical chapters, it is possible to group by both the choice set and the respondent (both of which are necessary, except in the conditional logit where all responses are independent). However, in the random effects models, the STATA command does not allow identification of both cluster variables. The use of *xtset* provides one, but the command offers neither the *id* option nor the *group* option. Therefore, the data was coded as differences, where one line represents one choice set, and each column represents the difference between the value of the parameter under option A minus the value of the parameter under option B. Thus, *xtset* was used to identify the respondent. The choice was either coded as 1 (if A was the preferred option) or -1 (if B was preferred). The impact of this is small, except that the log likelihoods between the random-effect models and the rest are not comparable, which therefore also applies to the Information Criteria.

$$U_{isj} = X'_{isj} \beta_i + \varepsilon_{isj} / \sigma \quad \text{Equation 35}$$

This utility function can be rewritten as

$$U_{isj} = X'_{isj} \beta \sigma_i + \varepsilon_{isj} \quad \text{Equation 36}$$

Model B is the scaled multinomial logit model (using Fiebig *et al*'s (2010) label, the S-MNL). Thus, relative to the conditional logit in Model 1, one additional parameter is estimated. Note that σ_i is not estimated for each respondent; this would lead to a huge number of additional coefficients. Rather, a distribution is estimated. As scale has to be constrained to be positive, the distribution is conventionally assumed to be log-normal

$$\ln \sigma_i \sim N(\bar{\sigma}, \tau) \quad \text{Equation 37}$$

STATA code, developed by Gu *et al.* (2011) and used in this thesis, normalises $\bar{\sigma}$ to be one, so the output reports the value of τ . The S-MNL is a highly parsimonious method for characterising heterogeneity relative to other methods outlined below as it only introduces one extra parameter. A recent paper has considered the use of latent class analysis of preferences regarding health (Flynn, et al., 2010); this was not considered in this analysis. While Flynn argued that a latent class approach has the advantage of not requiring parameterisation of heterogeneity, Hole (2008) identified that the conclusions from a latent class approach and a mixed logit give similar results, both representing significant improvements relative to a standard logit. Indeed, Keane and Wasi (2009) demonstrate using a range of datasets that latent class analysis is consistently outperformed by other approaches, including the S-MNL (and the G-MNL which is discussed below as Models 4 and 6).

However, while the S-MNL may offer substantial improvement on the conditional logit or random-effects probits or logits, scale heterogeneity is not the only possible source of heterogeneity. Preference heterogeneity, in which different attributes are of relatively greater or lesser importance to different respondents (independent of scale), may be an additional area in which the conditional logit is overly restrained.

Heterogeneity exploration models 3 and 5 - Mixed logit analysis

In the previous section, the importance of accounting for scale heterogeneity was outlined, and it was noted that other sources of heterogeneity exist and should be

considered. A model which does this is required, the most obvious of which is the mixed logit, which has become increasingly common in health economics (Hall, et al., 2006a; Hole, 2007a; Hole, 2008). This approach allows for possible heterogeneity among individuals by setting

$$\beta_{ki} = Z_i' \bar{\beta}_k + \sigma_k \omega_{ki}, \quad \text{Equation 38}$$

where Z_i and $\bar{\beta}_k$ are a vector of observed characteristics of a respondent i and a vector of parameters respectively, and $\sigma_k \omega_{ki}$ represents unobserved heterogeneity in the preference weights. Concern with modelling heterogeneity in preferences between respondents is longstanding (see for example Hausman and Wise (1978)), but solutions remained theoretical until the recent increase in computing processing capabilities. The ω_{ki} terms are usually assumed to follow standard normal distributions (although log-normal distributions can be imposed in situations where a positive coefficient is required), which are independent both of each other and of the error term in the utility function. This specification allows for the panel nature of the data as β_{ki} differs over individuals, but not over the repeated choices made by each individual.

One issue with the use of mixed logit is that it generates heteroskedastic error terms. The random utility model described previously becomes

$$U_{isj} = \mathbf{x}_{isj}' \beta_i + \varepsilon_{isj}, \beta_i = \beta + \eta_i \quad \text{Equation 39}$$

As discussed previously, the β_i term varies over individuals. Rather than allowing one parameter for each individual, it is assumed to be composed of a mean coefficient β and a variability term η_i which is distributed normally with zero mean and a non-zero standard deviation (Hildreth and Houck, 1968). The heteroskedastic error term results because the model is estimated based on the β term (the mean) rather than the β_i term. In notation,

$$U_{ij} = \mathbf{x}_{ij}' \beta + (\mathbf{x}_{ij}' \eta_i + \varepsilon_{ij}) \quad \text{Equation 40}$$

Thus, the error term is different based on the value of the \mathbf{x}_{ij}' .

A further issue with mixed logit models is that the vector of consumer utility weights for a particular level of an attribute is usually assumed to have a multivariate normal

distribution with mean zero and a vector of standard deviations Σ (Fiebig, et al., 2010). It should be noted that this is not necessarily the case: McFadden and Train highlighted that

“...any discrete choice model derived from random utility maximisation has choice probabilities that can be approximated as closely as one pleases by a (Mixed Logit)” (McFadden and Train, 2000) (p.447)

However, alternate distributions are rarely used (for example, the STATA command allows only a log-normal distribution as an alternative). Regarding the use of a multivariate normal distribution, it has been argued by Louviere (2008) that this distribution is not realistic, and that much of the differences in attributes between individuals are a result of a scale effect. As Flynn (2010) states, there is a perfect confounding of estimates of the mean and variance on the underlying latent scale. Conventionally, this variance scale is set to one to enable identification (thus not requiring explicit modelling of scale). The consequence of this is that a high coefficient on an attribute for an individual may represent the individual favouring that attribute, or being particularly certain in their preferences, or some combination of the two. Therefore, the distribution of a coefficient around a mean reflects both true heterogeneity in preferences, and also heterogeneity in certainty.

In the empirical chapters in this thesis, estimation by maximum simulated likelihood was undertaken in STATA, with all coefficients potentially varying across individuals. Choice probabilities are estimated in the following way:

$$P\left(j | X_{nt} = \frac{1}{D} \sum_{d=1}^D \frac{\exp[(\beta + \eta^d)x_{isj}]}{\sum_h \exp[(\beta + \eta^d)x_{ish}]} \right) \quad \text{Equation 41}$$

STATA takes D draws from the multivariate normal, and averages logit expressions over these draws to simulate choice probabilities. The mixed logit regressions are presented as Models 3 and 5. All mixed logit models were estimated using 500 Halton draws (as were those in the G-MNL, which is described below).

It is highly plausible that, in some situations, coefficients are likely to be correlated within an individual. Using a marketing example, individuals who prefer a pizza with ham may be more likely than average to favour a pizza with salami. In situations in which coefficients are likely to be correlated within an individual, the simple mixed

logit, in which all draws are independent, can be extended by relaxing the assumption of independent coefficients. Thus, while it is most conventional to assume that the η_i vector of coefficients is multivariate normal $(0, \Sigma)$, and that Σ is diagonal, this need not be so. Relaxing this assumption can be done in most statistical packages, such as by using the *corr* option in STATA. Thus, model 5 (and 6 discussed below) include this relaxation, exploring whether the relaxation adds predictive strength to the results of models 3 and 4 which make the limiting assumption. More details of this technique are given below in the outline of model 6 in the G-MNL.

Heterogeneity exploration models 4 and 6 - Generalised Multinomial Logit modelling

Recent literature has suggested that allowing for preference heterogeneity in the mixed logit without simultaneously allowing for scale heterogeneity impacts on the interpretability of the results (Fiebig, et al., 2010; Hensher, et al., 1999; Louviere, et al., 2008; Louviere, et al., 2002; Louviere, et al., 1999). The issue with considering the heterogeneity identified in mixed logit as preference heterogeneity is that, across individuals, all attribute weights can be scaled up and down in parallel due to the randomness or otherwise of their responses. This scale effect can be controlled for in a scale heterogeneity multinomial logit model. However, a step beyond this is to nest both preference heterogeneity and scale heterogeneity into one model, described by Fiebig *et al.* (2010) as a generalised multinomial logit model (G-MNL).

In this, the utility of choosing alternative j to an individual i in scenario s is given by

$$U_{ijs} = [\sigma_i \beta + \gamma \eta_i + (1 - \gamma) \sigma_i \eta_i] x_{ijs} + \varepsilon_{ijs}, \quad \text{Equation 42}$$

where γ is a parameter representing how the variance of residual taste heterogeneity varies with scale and σ_i is the scale parameter associated with individual i . As γ approaches 1, the scale term applies only to the β : conversely, as it approaches zero, the scale term is applied increasingly to the individual variability from the β : at the extreme, it applies equally to the parameter coefficient β and the variance term η_i .

Fiebig *et al.* (2010) define these two extreme cases as G-MNL-I and G-MNL-II respectively. Thus, G-MNL-I is

$$U_{ijs} = (\sigma_i \beta + \eta_i) x_{ijs} + \varepsilon_{ijs}, \quad \text{Equation 43}$$

while G-MNL-II (termed by Greene and Hensher (2011) as the scaled mixed logit model) is

$$U_{ijs} = (\sigma_i(\beta + \eta_i))x_{ijs} + \varepsilon_{ijs} \quad \text{Equation 44}$$

The code of Gu *et al.* (2011) employed in this thesis to estimate the G-MNL allows γ to be freely estimated; that is, it can take any value (Keane and Wasi, 2009). This moves away from the assumption in Fiebig *et al.* (2010) that estimates it in a 0-1 range. Keane and Wasi argue this is appropriate for two reasons. Firstly, constraining it in this way ignores the case in which the scale term applies more to η_i than to β . Secondly, the method used by Fiebig *et al.* estimated γ^* , rather than γ , where $\gamma = \exp(\gamma^*)/(1+\exp(\gamma^*))$; this caused problems as γ^* often tended to $\pm\infty$ (which would suggest that γ was close to 0 or 1). However, the interpretability of an unconstrained γ is not always clear. As Keane and Wasi (2009) note, if γ takes a negative value, it means that the scale term applies more to η_i than to β (which is plausible). This can be seen in Equation (42) as the term in the round brackets exceeds one, which is then multiplied by $\sigma_i\eta_i$. However, if γ takes a value greater than one, the term in the round brackets becomes negative, meaning a negative scale effect applies to η_i ; what this means is unclear and should be considered with suspicion. In the empirical chapters, the G-MNL is estimated using an unconstrained γ . If the result suggests $\gamma > 1$, then the model will be re-run constraining it to be 1 or less.

Interestingly, Fiebig *et al.* (2010) identify that, in the choice sets surrounding health questions specifically, the move from MNL to G-MNL leads to a relatively greater improvement in log-likelihood than in other, non-health, data sets. Indeed, the percentage improvement in log-likelihood was approximately twice as large in the health-based data sets as in the others. Flynn (2010) claims this pattern is not surprising:

“Variation in choice consistency is much lower when people are deciding which TV to buy than when they are choosing health insurance plan or treatment” (p.7)

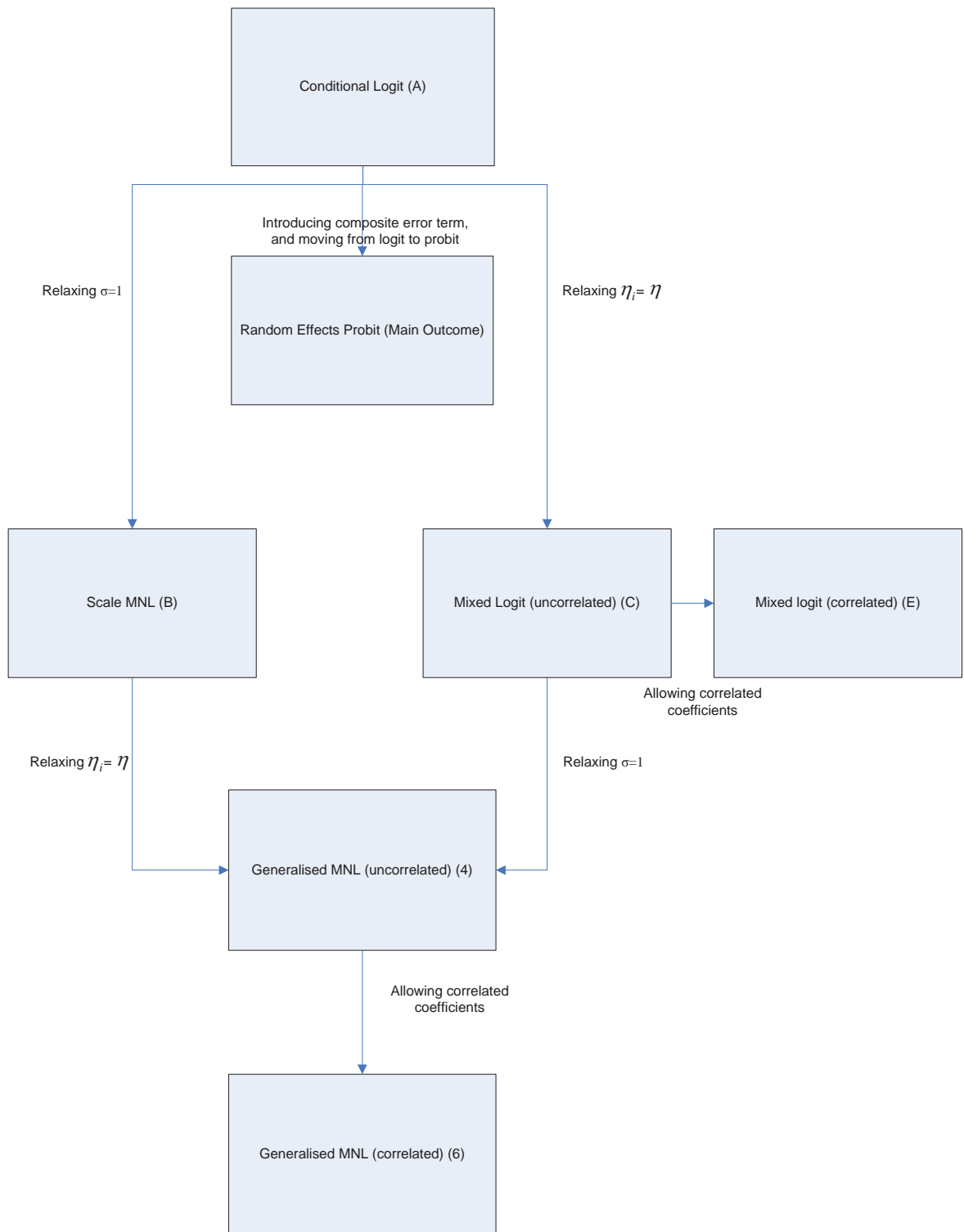
Other recent evidence has acknowledged the role of scale heterogeneity (Greene and Hensher, 2011). However, Greene and Hensher argue that, while accounting for scale heterogeneity might improve model fit, the importance of allowing for it in logit

models is of a much smaller magnitude that accounting for preference heterogeneity, and does not impact on estimates of (for example) elasticities or willingness to pay. The relative importance of accounting for the two sources of heterogeneity will be looked at empirically in our data sets.

Heterogeneity models 4 and 6 extended heterogeneity models 3 and 5 respectively, but replacing the mixed logit framework with this G-MNL one (the coding of models allows the number of parameters to increase from model 1 through to model 6). Gu *et al.* (2011) note that model convergence and time to convergence are highly sensitive to the starting value assumed by the analyst. In the empirical chapters that follow, a variety of approaches are taken, with the reported result noting the starting values employed. There are a variety of sensible starting values for the G-MNL. Using coefficients from the corresponding mixed logit model (correlated or not) is one possibility. As the mixed logit does not estimate either γ^* or τ , the starting value of these can be assumed (usually with $\tau = 1$ and $\gamma = 0.5$). In both empirical chapters in this thesis, the results for the G-MNL will specify which starting values achieved convergence.

It is important to note that all models estimated to this point are nested within the G-MNL. The models could therefore all be considered as G-MNL models with specific constraints placed on parts of the utility function. This point was made by Fiebig *et al.* (2010); how the modelling approaches incrementally build on each other is presented in Figure 11.

Figure 11: Nesting Regression Models Within the G-MNL



Two brief computational issues

With models employing simulated maximum likelihood, it is necessary to specify two major technical inputs to allow software to find the appropriate coefficients, these being the number of Halton draws and the method of optimization. The Halton draws used in the G-MNL estimation procedure were generated using the Mata function *halton()* (Drukker and Gates, 2006). Regarding the number of Halton draws, STATA defaults to 50. While this represents a low number of draws, and therefore convenient in terms of time to convergence, it has been argued that this may lead to serious convergence issues (Gu, et al., 2011). Train (2003) outlined the concept of Halton draws and discussed their use in preference to random draws. He cited Bhat (2001), who showed that 100 Halton draws for a mixed logit performed better (in terms of precision of estimates) than 1,000 random draws. However, whether these fewer Halton draws can be somehow defined to be precise enough is moot. The modelling in this thesis uses 500 Halton draws unless otherwise stated. With regard to optimization strategy, the default Newton-Raphson approach was taken, which is described in Chapter 8 in Train (2003). While estimation under Newton-Raphson is slow, my experience is that it is more likely to eventually reach convergence in simulated models.

Model evaluation

The major approach to model evaluation was the use of Akaike and Bayesian information criteria (AIC and BIC) (Akaike, 1974; Schwarz, 1978). These consider both the model fit and also the parsimony of the model (by accounting for the number of parameters in the model). AIC is estimated using the log likelihood and the number of parameters estimated:

$$AIC = 2k - 2 \log \text{likelihood} , \quad \text{Equation 45}$$

where k is the number of coefficients estimated. The BIC has a relatively greater emphasis on parsimony; therefore, disagreement concerning preferred specification (defined by minimising the coefficient) is possible between the two. BIC is estimated in the following way:

$$BIC = k \ln(n) - 2 * \log \text{likelihood} , \quad \text{Equation 46}$$

where n is the number of observations. There is however an issue with the use of BIC in panel data with multiple observations per respondent. It is unclear whether n should represent the number of clusters (in this case the number of respondents) or the number of observations (which, in a data set with only complete responders, is equal to the number of choice sets multiplied by the number of respondents). It depends on the degree to which observations within a respondent are independent: the more this statement is true, the more appropriate it is to use the number of observations rather than the number of respondents. In this case, it is unclear which approach is appropriate. Certainly, there will be considerable agreement between responses from one individual; however, to assume this agreement is perfect is very strong. Therefore, both BIC estimates are calculated, and any disagreement in ranking of models between them are discussed further.

Fiebig *et al.* (2010) investigate the relative merits of AIC and BIC for the models estimated here. They generate twenty simulated datasets for each of the S-MNL, the Mixed logit (correlated errors) and the G-MNL (correlated errors) for two contexts (pap smear tests and holidays). They then run multinomial logit, S-MNL, Mixed Logit and G-MNL on the simulated data and identify the probability of each Information Criteria identifying the correct underlying utility function. When the true utility model is S-MNL, all Information Criteria have perfect prediction (in that the preferable model in information criteria terms was the S-MNL). When the true utility function corresponded with the Mixed Logit, the AIC identified it correctly in approximately half of the simulations; otherwise, it suggested the G-MNL was preferred. The BIC performed less well. In no simulation did it identify correlated errors, although it did tend to prefer the mixed logit with uncorrelated errors to the G-MNL with uncorrelated errors. If the true model was G-MNL, the AIC performed well identifying it as such every time, but preferring the more parsimonious uncorrelated error specification in eleven of the twenty simulations in the Pap Smear context. Arguably, this suggests the AIC is a better method of identifying the underlying utility function.

Deriving welfare measures from discrete choice experiments

Up to this point, the chapter has discussed issues in the conception (as opposed to the design), analysis and evaluation of discrete choice data. One issue which has not been

discussed is the appropriate grouping of profiles in surveys. This is the major focus of Chapter 4. Before doing this, it is necessary to discuss how results from a choice experiment can best be made policy relevant. The issue is that identifying whether A is preferred to B (where A and B can be products, or levels of a dimension in the choice experiment) is valuable, but it is necessary to know how much better A is than B, and to put it in a metric which can be used in policy. It has been widely acknowledged in utility theory and in the psychology literature that comparisons of coefficients between experiments to identify relative importance is fundamentally flawed (Keeney and Raiffa, 1993; Lancsar, et al., 2007).

Lancsar *et al.* (2007) present some options for evaluating the relative importance of attributes in stated preference experiments. Of most relevance to this work are (i) Hicksian welfare measures such as compensating or equivalent variation (CV / EV); and (ii). marginal rates of substitution (*MRS*)

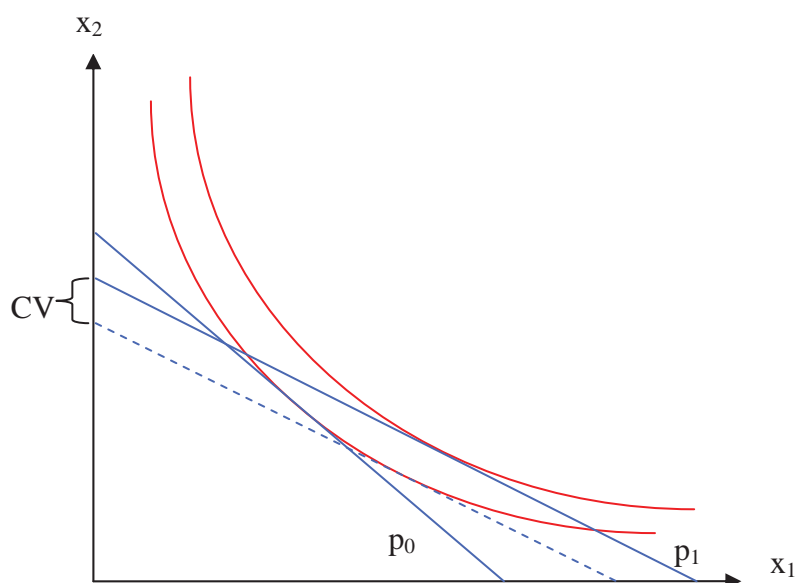
The Hicksian Compensating Variation

Using the Hicksian compensating variation (CV) to explore the relative importance of attributes within a choice experiment was first discussed by Small and Rosen (1981), and then introduced to the Health Economics literature by Lancsar and Savage (2004). The CV, when considering the willingness to pay for some policy change, is estimated as

$$CV = \frac{1}{\lambda} \left[\ln \sum_{j=1}^J e^{V_j^0} - \ln \sum_{j=1}^J e^{V_j^1} \right], \quad \text{Equation 47}$$

where λ is the marginal utility of income, V_j^0 and V_j^1 are the values of the IUF for each choice option j before and after the policy change, and J is the number of options in the choice set. The CV is illustrated diagrammatically in Figure 12.

Figure 12: The Compensating Variation (CV)



In a hypothetical market, there are two goods, x_1 and x_2 . The initial budget constraint (the blue line marked p_0) demonstrates the relative price of the two goods. The red indifference curves reflect combinations of the two goods that have equal utility for the consumer. Imagine that the relative price of the two goods changes from p_0 to p_1 . Thus, x_1 becomes relatively and absolutely less expensive, and the consumer can move to the higher indifference curve. The CV is defined as the income change required to compensate the consumer for the price change (Varian, 1984) and uses the new prices as the base (thus differentiating it from the EV which uses the original prices). In this case, the CV is negative as the consumer is on a higher indifference curve following the price change.

The foundation of the CV technique (and the equivalent variation) lies in the work of Hicks (1939), and was adapted for the discrete situation by Small and Rosen (1981). The compatibility of the CV technique with RUT has been widely discussed, and in the health economics context by Lancsar and Savage (2004).

The advantage of this approach is that it accounts for the probability that each alternative will be chosen by a typical respondent. Consider a choice set with eight options A-I, with some initial ranking of the options with A most likely to be selected, down to I being least likely. If a policy improves I but not to the extent that I is

significantly more likely to be picked, the willingness of the typical respondent to pay for this improvement in I is likely to be effectively zero.

Marginal rates of substitution

Using marginal rates of substitution as a method for deriving a welfare measure for the utility associated with a changing attribute is easy to apply and an intuitive method. There are a large number of applications of the approach in the health economics literature (Gyrd-Hansen and Sogaard, 2001; McIntosh and Ryan, 2002; Scott, 2001). The *MRS* is calculated by partially differentiating an indirect utility function V with respect to one attribute x_1 , and then with respect to another attribute x_2 , then calculating a ratio, i.e.,

$$MRS_{x_1, x_2} = \frac{\delta V / \delta X_1}{\delta V / \delta X_2}. \quad \text{Equation 48}$$

Thus, the numerator is the marginal utility of X_1 , and the denominator is the marginal utility of X_2 . Using the ratio puts the marginal utility of X_1 in the units of X_2 . If X_2 is a price, the *MRS* represents a marginal willingness to pay for a change in X_1 . In a main effects model, this term is usually interpreted as a ratio of coefficients (although Lancsar *et al.* (2007) have shown that Equation (48) is a more general expression which continues to be applicable under different specifications of the utility function).

There has been considerable discussion in the literature regarding the relative appropriateness of CV and *MRS* (Ryan, 2004; Santos Silva, 2004). Ryan provides an interesting distinction illustrating the types of situations in which the two approaches are preferred. She distinguishes between ‘state-of-the-world models’ and ‘multiple alternative models’. In the former, there is “only one alternative on offer at any one time, and individuals take up the service / drug with certainty.” (p.909). If this is true,

she argues that CV reduces to *MRS* as $\ln \sum_{j=1}^J e^{V_j^0}$ reduces to V_j^0 . Importantly, Ryan

then argues that ‘state-of-the-world models’ are more likely in health than in, for example, transport or environmental issues as choices are often limited. The caveat to this point is that non-demanders are important to model. In the CV, if the options before and after the policy change are both unattractive, and the potential patient is able to opt out of treatment (which is normally the case), the CV becomes

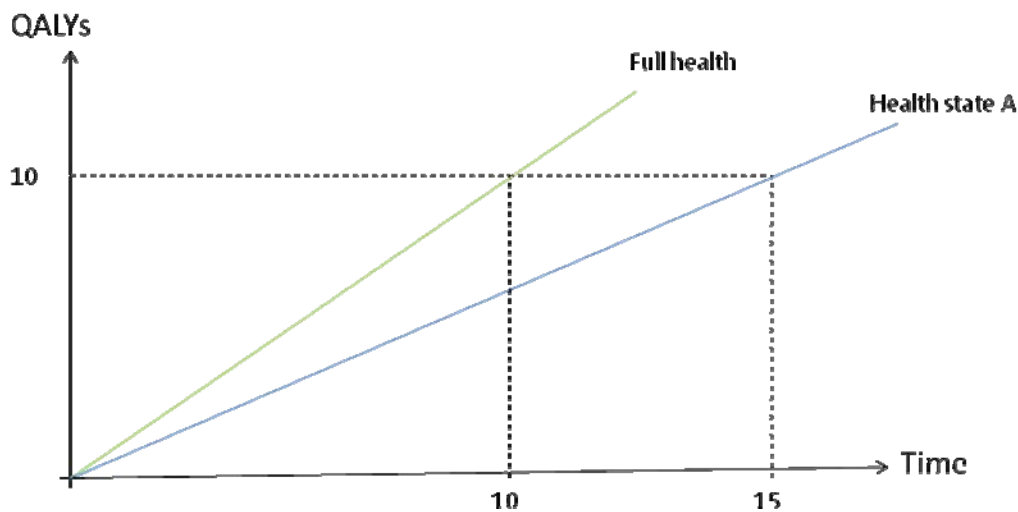
$$CV = \frac{1}{\lambda} \left[\left(e^0 + \ln \sum_{j=1}^J e^{V_j^0} \right) - \left(e^0 + \ln \sum_{j=1}^J e^{V_j^1} \right) \right], \quad \text{Equation 49}$$

which moves further from the CV stated in Equation (47) as the V_j^0 and V_j^1 terms become increasingly large and negative (as the opt-out option becomes increasing likely to be selected even if the active options improve).

Using a ratio of marginal utilities

This can be adapted to the framework established in Equations (23) and (25). The need for adaptation from the conventional *MRS* approach is that, in the empirical Chapters 5 and 6, I am interested in the value of a health profile relative to some other health profile (e.g. the value of quality of life in a particular health state relative to full health). This differs from *MRS* in that it is a comparison of two entire health states, rather than a comparison of two coefficients within one regression. Therefore, what is needed is the ratio of marginal utilities (*RMU*) associated with the two profiles. The reason for this can be illustrated diagrammatically, as in Figure 13.

Figure 13: Ratio of Marginal Utilities



In Figure 13, the approach underpinning the valuation of health profiles is presented using a standard QALY model. In this case, 10 years in full health by definition provides 10 QALYs. If some other health state A provides 10 QALYs over a 15 year period (and hence the overall profiles of health are valued equally), the QALY model

values that health state at $10/15 = 0.667$. This calculation is simply the ratio of the slopes of the two lines, and is the approach taken here.

This is derived from Equation (23) for the case in which utility is assumed to be linear with respect to time, and from Equation (25) if that assumption is relaxed by introducing quadratic terms. If it is assumed that the simpler utility function in Equation (23) is used, the *RMU* between two alternatives health profiles j and j^* can therefore be estimated as

$$RMU_{j,j^*} = \frac{\alpha + \beta X'_{isj}}{\alpha + \beta X'_{isj^*}}. \quad \text{Equation 50}$$

If the less constrained approach is taken, as exemplified by Equation (25), the *RMU* between the same two alternatives j and j^* can be estimated as

$$RMU_{j,j^*} = \frac{\alpha + \beta X'_{isj} + 2TIME_{isj}(\rho + \phi X'_{isj})}{\alpha + \beta X'_{isj^*} + 2TIME_{isj^*}(\rho + \phi X'_{isj^*})}. \quad \text{Equation 51}$$

This second estimate of *RMU* is clearly more difficult to estimate as it is now dependent on the value of *TIME*; while this may better represent the trade-offs that people make, a question that is raised in the following empirical chapters is whether accounting for different *RMU* is worthwhile when welfare measures are constructed.

In the empirical chapters that follow, I will use the *RMU* to estimate welfare measures; however, where a ‘multiple alternative model’ is plausible, I will also analyse the data using a CV approach and compare the results.

Chapter summary

This chapter has presented the methods used in the empirical chapters that follow it. Specifically, the discrete choice experiment has been described, beginning with the principle of probabilistic choices. The subsequent key issues raised were the choices surrounding the design of the choice experiment, the methods for allowing for and investigating response heterogeneity, and the appropriate specification of the utility function to allow for the unusual complementary relationship between time and other characteristics which are meaningless (and without value) in the absence of time.

Finally, the methods for deriving welfare measures were discussed. Using a ratio of

marginal effects was proposed, an approach which will be pursued in Chapters 5 and 6.

Chapter 4: Some Principles for Designing Discrete Choice Experiments

Chapter summary

Chapter 3 introduced the DCE as a potentially highly attractive tool to explore preferences in health. However, the description did not consider how to construct a discrete choice survey. This issue is of paramount importance as a failure to design the experiment to allow investigation of the parameters of interest renders the method of analysis irrelevant. In this chapter, the methods for designing statistically optimal DCEs are outlined, alongside some ongoing controversies in the area. The multifaceted nature of choices in the health field make designing small experiments difficult without considerable constraints on the effects that can be separately identified; thus, it is important to maximise the information the analyst derives from each question, and from each survey respondent, particularly in contexts in which the population of interest is limited in size. The concepts of contrasts and of arrays are introduced and discussed, particularly how the choices made in constructing arrays impact on what effects can be investigated. Different strategies for designing and evaluating choice experiments are then described. In particular, the chapter looks at the use of shifted designs and Kuhfeld's SAS algorithms, as they are the two methods employed in the empirical chapters of this thesis.

DCE design principles

This thesis presents two major pieces of empirical work in Chapters 5 and 6. Both are DCEs, and each uses a different approach to the design of the experiment. Chapter 5 uses SAS algorithms described by Zwerina, Huber and Kuhfeld (2010), while Chapter 6 uses a generator-type approach as suggested by Bunch, Louviere and Anderson (1996), and extensively developed in Street and Burgess (2007). After a brief introduction to design theory and some preliminary discussion, the two approaches will be outlined.

Introduction to design theory

Chapter 3 discussed the appropriate tools for analysing DCE data in the context of an experiment with fundamental complementarity of attributes, and the necessity of including a suitable numeraire when designing the experiment. It then discussed the

appropriate tools for estimating welfare measures from the results of a choice experiment. This chapter deals with a particular interpretation of design in DCEs, specifically the methods for selecting the pairs of profiles that will be shown to the respondents in the studies described in the empirical chapters.

In choice experiments, and also in regression analysis more generally, we are interested in investigating the impact of an explanatory variable (often called treatments) on some response variable in a set of experimental units. There are considerable advantages to being able to control the allocation of the explanatory variable between the units. In a comparative experiment (of which DCEs are an example), the experimenter chooses which treatments are administered to which experimental units. This compares with the more general observational studies in which

“the objective is to elucidate cause and effect relationships... (where) it is not feasible to... assign subjects at random to different procedures (or treatments)”. (Cochran and Chambers, 1965)

The need to account for differences in characteristics of experimental units prior to receiving the treatment is a perennial problem which comparative experiments seek to circumvent, as they are used in situations where the experimenter has control over the administration of treatments to units.

In a discrete choice experiment, exploring the impact of changing individual attributes one at a time on the choices of the respondent is an inefficient way of exploring preferences, and indeed may miss interactions altogether. As a result of considering the study of crop yields in agriculture, Sir Ronald Fisher introduced the idea of factorial designs which allow the simultaneous investigation of multiple factors (Fisher, 1935; Fisher, 1925). If we have k variables we are interested in (which will be called attributes), and each attribute has l_q possible values (called levels), the full factorial in which every possible combination is investigated requires l_q^k separate tests, a number which can easily prove unmanageable. More generally, the number of possible combinations L can be defined as $L = \prod_{q=1}^k l_q$ where there are k dimensions and the q th dimension has l_q levels. The principles of designed experiments address how to select a subset of this full factorial such that a set of effects can be estimated in

such a way that it optimises some pre-specified criterion, perhaps minimising variance. Burgess, Street and Wasi (In Press) summarise the problem,

“Which alternatives should we show together in choice sets, and which choice sets should we include in a discrete choice experiment?”

The optimal DCE in a given setting is a function both of some simple universal principles of design (e.g. there is no sensible information derived from making a respondent choose between identical options) and the kinds of issues the analyst wishes to investigate.

What are we trying to do when we design a choice experiment? Louviere *et al.* (2000) state four objectives,

1. Identification: The form of the utility function can be estimated from the experiment;
2. Precision: The parameters can be estimated as precisely as possible;
3. Cognitive complexity: The experiment should not be so difficult for the respondent that responses are adversely affected; and
4. Market realism: The choices should reflect as closely as possible the choices that people actually make.

Viney *et al.* (2005) argue that the focus has been on the first two of these, at the expense of the latter two. A recurring theme is that statistical efficiency, associated with the first two of these objectives, is often in conflict with the final two (which impacts on what can be termed *respondent efficiency*).

It should be noted that the fourth objective is of particular difficulty in a health setting. As noted in the introductory chapter of this thesis, the health sector involves significant government intervention and proxying of a market. Thus, the choice sets faced by an individual are often unavoidably unrealistic. One obvious instance of this is if the options include items such as cost; something which is needed if the analyst wishes to estimate willingness to pay for a service, but not something which healthcare consumers face in many of their real-world decisions.

The issue of respondent efficiency is necessarily context-specific. For example, in the context of the multi-attribute utility instruments described in Chapter 2, a DCE which investigates these kinds of issues ought to consider which combinations of quality of life are so implausible that the respondent would be likely to find difficulty answering

the question. Therefore, I will return to it in the empirical chapters which follow. The concept of statistical efficiency is more universal, so I will turn to this now. This includes both methods for constructing DCEs, and for evaluating how well the resultant designs achieve a set of pre-specified positive characteristics.

Contrasts and fractional factorial designs

An important concept requiring introduction at this point is that of contrasts. I will first demonstrate the principle of contrasts in a non-DCE setting. Consider a situation in which there are 3 attributes (called A , B and C) that might have some impact on a dependent variable (Y). Using an agricultural example, A , B and C might be use of *pesticide*, *provision of water* and *exposure to sunlight* respectively, and Y might be the crop yield from a plot of land. Initially, it might be assumed that each of A , B and C has only two levels (each labelled as 0 and 1, relating to appropriate low and high levels for example). Thus, there are eight possible treatments, which will be labelled lexicographically as 000, 001, 010, 011, 100, 101, 110 and 111 where the first figure in each treatment refers to the level of A , the second to the level of B , and the third to the level of C . Initially, the focus is on the situation in which all level combinations appear; this will be called a full factorial.

In an experiment, it is necessary to identify the effect of each parameter in the model independently (called the main effect of the parameter). If we look first at the main effect of A on Y (using the agricultural example, the impact of *pesticide* on yield), it is plausible that the effect depends on the levels of B and C (i.e. there are interactions). We can identify the simple effect of A and Y , defined as the impact of A for each of the possible combinations of B and C .

Thus, the simple effect of A if B and C are both set at 0 is

$$y_{100} - y_{000} \quad \text{Equation 52}$$

The main effect of A is then estimated as the mean of the simple effects over the four possible combinations of B and C , i.e.

$$((y_{100} - y_{000}) + (y_{101} - y_{001}) + (y_{110} - y_{010}) + (y_{111} - y_{011}))/4 \quad \text{Equation 53}$$

53

Street and Burgess (2007) have reported this result for 2^k designs:

$$\frac{1}{2^{k-1}} \left(\sum_{x_q=1} y_{x_1 x_2 \dots x_k} - \sum_{x_q=0} y_{x_1 x_2 \dots x_k} \right), \quad \text{Equation 54}$$

where k is the number of binary dimensions, and x_q is the value taken by the q th dimension (either 0 or 1). Assuming there are no interaction effects, each of the four terms in the numerator of Equation (53) have a common expected value, so averaging over them has the effect of reducing random variability. It is this assumption of a common expected value that is being tested when we begin looking for interaction effects.

A contrast is a linear combination of the responses with coefficients that sum to 0. Thus, it would be sensible to define a contrast to estimate the main effect of A (and for B and C , which are constructed similarly) in the following way:

Table 9: Contrasts for main effects in a 2^3 experiment

000	001	010	011	100	101	110	111	
-1	-1	-1	-1	1	1	1	1	A
-1	-1	1	1	-1	-1	1	1	B
-1	1	-1	1	-1	1	-1	1	C

In Table 9, a 1 in a cell means that the attribute of interest (A , B or C) has a value of 1 in the respective column. Correspondingly, a -1 means that the attribute has a value of 0 and would be subtracted when estimating the effect. Thus, A is 1 for 100, 101, 110 and 111, but -1 for 000, 001, 010 and 011.

For sets of contrasts, it is important that they are orthogonal to one another. This property ensures that the effects of these two levels can be estimated separately. Two contrasts are independent or orthogonal if the sum of the product of corresponding

coefficients is 0, i.e. that $\sum_t \frac{a_t b_t}{n_t} = 0$, where the product in the numerator is the

product of the figures for each of the treatments in Table 9 and the denominator is the number of observations for each treatment (Dey, 1985; Plackett, 1946). In Table 9, it is apparent that this condition is met (consider the inner product of each of the three pairs of rows).

The agricultural example used to this point is highly constrained in that binary attributes are not always appropriate, and interactions may also be of interest. I will return to the situation of attributes with more than two levels at a later point, and will

now focus on situations in which interactions are of interest. In the example I have just been discussing, it may be plausible that the effect of A on Y would depend on the levels of other attributes (in this case, B and C). In the agricultural example, this is highly plausible; for example, it might be that pesticides only impact on crop yield if the crop receives enough water. This would mean that we would expect certain of the simple effects to differ from the main effect through more than random response variability. In the case of A interacting with B , the interaction would be simple effect of A when B is at level 1 minus the simple effect of A when B is at level 0. If there is no interaction (and B does not impact on the effect of A), this difference will be expected to equal 0:

$$((y_{110} - y_{010}) + (y_{111} - y_{011}) - ((y_{100} - y_{000}) + (y_{101} - y_{001}))) / 2 \quad \text{Equation 55}$$

Table 9 can be expanded into Table 10 to give the contrasts that apply to both the three two-factor interactions AB , AC and BC , and the one three-factor interaction ABC

Table 10: Contrasts for main effects and interactions in a 2^3 experiment

000	001	010	011	100	101	110	111	
-1	-1	-1	-1	1	1	1	1	A
-1	-1	1	1	-1	-1	1	1	B
-1	1	-1	1	-1	1	-1	1	C
1	1	-1	-1	-1	-1	1	1	AB
1	-1	1	-1	-1	1	-1	1	AC
1	-1	-1	1	1	-1	-1	1	BC
-1	1	1	-1	1	-1	-1	1	ABC

Thus, the contrast on A is the sum of all of the combinations of A , B and C in which A is 1 minus the sum of all of the combinations of A , B and C in which A is 0. This can be seen in the second row of Table 10.

The contrasts for the interaction terms can be derived from first principles as above. This is shown in Equation (55), as the positive terms are reflected by 1s in Table 10, while the negative terms are reflected by 0s. Alternatively, the contrast for interactions can be derived by multiplying terms in the contrasts for the constituent parts of the corresponding interaction (in this case, two or three of the contrasts A , B or C). As before, these contrasts can be shown to be orthogonal to one another, and to the mean vector.

If the experiment requires additional levels within attributes, constructing contrasts requires additional thought as orthogonality is less easily maintained. Initially, it is more convenient to consider this with only two attributes, so I will remove C . So, now, suppose that A and B each have three levels (this time labelled 0, 1 and 2). It might appear logical to present contrasts in the following way:

Table 11: (Non-orthogonal) contrasts for main effects in a 3^3 experiment

00	01	02	10	11	12	20	21	22	
-1/2	-1/2	-1/2	1	1	1	-1/2	-1/2	-1/2	A_1
-1/2	-1/2	-1/2	-1/2	-1/2	-1/2	1	1	1	A_2
-1/2	1	-1/2	-1/2	1	-1/2	-1/2	1	-1/2	B_1
-1/2	-1/2	1	-1/2	-1/2	1	-1/2	-1/2	1	B_2

Note that there are an infinite number of ways of expressing the same contrast matrix by multiplying each term by a constant. It is conventional to fix a term at 1; however the conclusions drawn are independent of which term is fixed.

Under this approach, the contrasts would represent the move from the base level to each of the other two levels in both dimensions. The sum of the weights across each proposed contrast is 0. Additionally, the previously stated constraint for orthogonal

contrasts that $\sum_i \frac{d_i x_i}{n_i} = 0$ holds for some pairs of contrasts (those between either of the

A terms and either of the B terms). However, the pairs of contrasts A_1 and A_2 , and B_1 and B_2 are clearly not orthogonal, and hence are not independently estimable under this approach. The conventional approach to dealing with this is to separate each term into a linear and quadratic term (for three-level attributes; more generally, for n -level attributes, $n-1$ terms are required representing the linear, quadratic, quartic etc). The contrasts in this situation, extended to consider interactions in the same way as in the binary attribute case, are presented in Table 12, which is based on Table 2.2 of Street and Burgess (2007).

Table 12: A, B, and AB contrasts for main effects and interactions in a 3³ experiment

00	01	02	10	11	12	20	21	22	Tmt. combination
1	1	1	1	1	1	1	1	1	Mean
-1	-1	-1	0	0	0	1	1	1	A Linear = A _L
1	1	1	-2	-2	-2	1	1	1	A Quadratic = A _Q
-1	0	1	-1	0	1	-1	0	1	B Linear = B _L
1	-2	1	1	-2	1	1	-2	1	B Quadratic = B _Q
1	0	-1	0	0	0	-1	0	1	A _L x B _L
-1	2	-1	0	0	0	1	-2	1	A _L x B _Q
-1	0	1	2	0	-2	-1	0	1	A _Q x B _L
1	-2	1	-2	4	-2	1	-2	1	A _Q x B _Q

This approach remedies the issue of non-orthogonality as, for each pair of contrasts,

$$\sum_t \frac{A_{ij} B_{ij}}{n_i} = 0.$$

A full factorial can become unmanageable in terms of respondent burden, and the time and money required to collect data. An approach to deal with this problem is to apply a fractional design in which a subset of the full factorial is selected so that it is possible to estimate pre-specified parameters of interest. The downside of using a fractional factorial is that certain effects cannot be separately estimated. However, Montgomery (2005) (in an Ordinary Least Squares context) describes an example full factorial, and the types of effects that are estimable,

“(A) complete replicate of (a) 2⁶ design requires 64 runs. In this design, only 6 of the 63 degrees of freedom correspond to main effects, and only 15 degrees of freedom correspond to two-factor interactions. The remaining 42 degrees of freedom are associated with three-factor and higher interactions” (p.282)

There are many settings in which interactions involving many factors are difficult to interpret. Hence it is not necessary to think about estimating them when designing an experiment. Thus a smaller design can be used to investigate the effects which are interpretable. I will outline this idea in the context of design more generally (i.e. without combination of options into choice sets), before summarising how the approach can be extended to a DCE setting. A further preliminary point to make is

that I will focus on regular fractions; for a discussion on irregular fractions, see Addelman (1961), or Chapter 2 of Street and Burgess (2007).

The concept of regular fractions will first be described in the simpler two-level situation, and then extended to higher numbers of levels in each dimension. Street and Burgess (2007) define a regular fraction of a 2^k factorial as

“a fraction in which the treatment combinations can be described by the solution to a set of binary equations... The binary equations are called the defining equations or the defining contrasts of the fractional factorial design. A regular 2^{k-p} fraction is defined by p independent binary equations...”(p.27).

Thus, in a 2^k setting, each binary equation will halve the size of the fraction. The largest fraction is obviously one with only a single binary equation, which will have half the runs of the full factorial. If $k = 5$ (for example), a common approach to produce a 2^{5-1} design is to take the full 2^4 design (i.e. each of the 16 combinations of four binary variables) and define the fifth variable as the product of the existing four (where each of the four binary variables takes either the value -1 or 1). This approach is used to produce column *E* in Table 13.

Table 13: A 2^{5-1} fractional factorial design

A	B	C	D	E
-1	-1	-1	-1	1
-1	-1	-1	1	-1
-1	-1	1	-1	-1
-1	-1	1	1	1
-1	1	-1	-1	-1
-1	1	-1	1	1
-1	1	1	-1	1
-1	1	1	1	-1
1	-1	-1	-1	-1
1	-1	-1	1	1
1	-1	1	-1	1
1	-1	1	1	-1
1	1	-1	-1	1
1	1	-1	1	-1
1	1	1	-1	-1
1	1	1	1	1

Note that the generation of the figures in column E can be described additively rather than multiplicatively. In this case, the -1s become 0s, and then, under addition modulo 2 (so $1+1=0$), $E=-(A+B+C+D)$.

Box, Hunter and Hunter (1978) note this solution to generating fractional factorials, and ask whether this solution limits the analyst in investigating effects. The answer is yes, in that the reduction in the number of runs implies certain interactions are confounded. For example, $ABC = DE$ (a fact which can be confirmed in Table 13 by multiplying the terms and comparing), This can also be described as these effects being *aliased*. If a different approach were used for constructing a fractional factorial, then different pairs of effects would be confounded. The question the analyst has to ask when designing experiments is which set of pairs of effects are more important (and must not be confounded), and whether the loss of separate estimability of these terms is a significant enough loss to the experiment to justify use of the full factorial (if there are enough experimental units to allow this approach).

The choice of binary equations is important as they determine which effects can be estimated independently within the fraction. If we have $k = 5$, and require a design in 8 runs (so, 2^3 or 2^{5-2} runs), the following equations might be plausible binary equations.

$$x_1 + x_2 + x_3 + x_4 = 0$$

$$x_1 + x_3 + x_5 = 0$$

By summing these, we also know that $x_2 + x_4 + x_5 = 0$. Thus, two binary equations produce three constraints (although only two of the constraints are independent as summing any two produces the third). Note that the array produced when these equations are set to 0 is known as the *principal fraction*. If there are two equations, as in the example here, that means there are 2^2 possible fractions which are mutually exclusive and exhaustive, reflecting that each fraction represents $1/2^2$ of the full factorial. The values in the fraction defined by this pair of equations are (0,0,0,0,0), (0,0,1,1,1), (0,1,0,1,0), (0,1,1,0,1), (1,0,0,1,1), (1,0,1,0,0), (1,1,0,0,1) and (1,1,1,1,0). The four fractions under these binary equations are represented in Table Table 14.

Table 14: Non-overlapping regular designs

$x_1 + x_2 + x_3 + x_4 = 0$	$x_1 + x_2 + x_3 + x_4 = 0$	$x_1 + x_2 + x_3 + x_4 = 1$	$x_1 + x_2 + x_3 + x_4 = 1$
$x_1 + x_3 + x_5 = 0$	$x_1 + x_3 + x_5 = 1$	$x_1 + x_3 + x_5 = 0$	$x_1 + x_3 + x_5 = 1$
0 0 0 0 0	0 0 0 0 1	0 0 0 1 0	0 0 0 1 1
0 0 1 1 1	0 0 1 1 0	0 0 1 0 1	0 0 1 0 0
0 1 0 1 0	0 1 0 1 1	0 1 0 0 0	0 1 0 0 1
0 1 1 0 1	0 1 1 0 0	1 0 0 0 1	1 0 0 0 0
1 0 0 1 1	1 0 0 1 0	0 1 1 1 1	0 1 1 1 0
1 0 1 0 0	1 0 1 0 1	1 0 1 1 0	1 0 1 1 1
1 1 0 0 1	1 1 0 0 0	1 1 0 1 1	1 1 0 1 0
1 1 1 1 0	1 1 1 1 1	1 1 1 0 0	1 1 1 0 1

The question is then what information can we derive from each set of 8 runs in the fractional factorial. This can be demonstrated by identifying the effects that are aliased. For the case where the two binary equations are set to 0 (i.e. the left-hand column), the aliasing structure can be illustrated in the following way. We know that

$$x_1 + x_2 + x_3 + x_4 = x_1 + x_3 + x_5 = x_2 + x_4 + x_5 = 0.$$

Therefore, as we are working in modulo 2, it follows that we can identify the effects that are confounded with main effects, i.e.,

$$\begin{aligned} x_1 &= x_2 + x_3 + x_4 = x_3 + x_5 = x_1 + x_2 + x_4 + x_5 \\ x_2 &= x_1 + x_3 + x_4 = x_4 + x_5 = x_1 + x_2 + x_3 + x_5 \\ x_3 &= x_1 + x_3 + x_4 = x_1 + x_5 = x_2 + x_3 + x_4 + x_5 \\ x_4 &= x_1 + x_2 + x_3 = x_1 + x_3 + x_4 + x_5 = x_2 + x_5 \\ x_5 &= x_1 + x_2 + x_3 + x_4 + x_5 = x_1 + x_3 = x_2 + x_4 \end{aligned}$$

These relationships follow because, within each of the three constraints, each set of n terms must be equal to the remaining $(t-n)$ terms in that constraint where t is the total number of terms in the constraint. Thus, it is clear that these sets of effects cannot be separately estimated under this fraction. The main effect of A is confounded with CE , with BCD and $ABDE$. It is noteworthy that, of the 10 two-factor interactions, only six are confounded with a main effect. The ones that are not confounded with a main effect are AB , AD , BC and CD . If the analyst has a particular *a priori* reason for believing that certain interactions are important, then it may be possible to choose the defining equations so that these effects are not confounded with main effects.

However, in this case, these two factor interactions will be confounded with higher-order interactions i.e.,

$$x_1 + x_2 = x_3 + x_4 = x_2 + x_3 + x_5 = x_1 + x_4 + x_5$$

$$x_1 + x_4 = x_2 + x_3 = x_3 + x_4 + x_5 = x_1 + x_2 + x_5$$

The degree to which the choice of fraction impacts on the ability to separately estimate effects is usually defined in terms of resolution. If no main effect is confounded with another main effect, then the design at least resolution 3 (often called strength 2, where strength is always resolution minus one). A design which has resolution 3 is sometimes called an orthogonal main effects plan (OMEP). If a design is resolution 3, it means that at least one main effect is confounded with a two-factor interaction term. Similarly, a design of resolution 4 does not have any main effect confounded with either another main effect or a two-factor interaction, but there is at least one main effect which is confounded with at least one three-factor interaction. Therefore, each of the $\frac{1}{4}$ fractions described in Table 14 are resolution 3 as at least one (indeed all five) main effects are confounded with at least one two-factor interaction.

Box, Hunter and Hunter (1978) note that the resolution of an array can be calculated directly from the defining relation,

“(A) design of resolution R is one in which no p-factor effect is confounded with any other effect containing less than R – p factors... In general, the resolution of a two-level fractional design is the length of the shortest word in the defining relation.” (p.385)

Webb (1968) provides an alternative specification for resolution which is generalisable to both regular and irregular fractions,

“A fractional design is of resolution (2R+1) if it permits the estimation of all effects up to R-factor interactions, when all effects involving (R+1) factors and higher-order effects are assumed to be negligible.”

Street and Burgess (2007) provide tables (2.9 and 2.10) identifying the smallest known 2-level designs of at least resolution 3 and at least 5 (p.32).

The use of generators or equations can be extended to an l^k situation when $l > 2$. As with the binary case, the impact of additional ternary equations (i.e. where $l=3$) is to cut the full factorial into 3 (with two equations partitioning the complete factorial into 9 fractions, three independent equations partitioning the complete factorial into 27 etc). The contrasts often reflect the linear, quadratic etc terms as discussed previously, although other subdivisions are possible and may be more appropriate. As with the binary case, Street and Burgess (2007) report the smallest known regular 3-level designs with resolution of at least 3 (Table 2.16) and of at least 5 (Table 2.17). Good discussions of this can be found in all major reference books in the area (Box, et al., 1978; Montgomery, 2005; Street and Burgess, 2007).

In practice, there are existing libraries of orthogonal arrays available online which the practitioner can adapt to their own purposes. Good examples of these are the libraries maintained by Sloane (<http://www.research.att.com/~njas/oadir/index.html>) and Kuhfeld (http://support.sas.com/techsup/technote/ts723_Designs.txt). The Kuhfeld library is limited to designs of resolution 3, so are not suited to situations in which interactions are to be investigated. Street and Burgess (2007) provide some strategies for adapting existing arrays to a different setting (e.g. with more dimensions, more levels), such as collapsing of levels, expansive or contractive replacement, adding factors or juxtaposing two orthogonal arrays (pp.46-51).

Using the runs of an orthogonal array to define one option in a choice set is a standard approach (and one which will be followed in Chapter 6 of this thesis). How to generate second and subsequent options in each choice set is the focus of the section later in this chapter looking at shift generators. However, before looking at this, I will discuss strategies for comparing different designed experiments.

The likelihood function and maximum likelihood estimators

Previously, it was noted that, in the context of the multinomial logit (MNL), the probability of individual i selecting option j in choice set s was assumed to be

$$P_{isj} = \frac{\exp(X'_{isj} \beta_i)}{\sum \exp(X'_{ish} \beta_i)} \quad \text{Equation 56}$$

For simplicity, $\pi_i = e^{V_i}$, where V_i is the systematic component of the utility function.

I will now describe the likelihood function and the principle of maximum likelihood estimation (MLE) which is a key step in comparing designs. The concepts will be described within a Bradley-Terry model in which all choice sets are pairs (Bradley and Terry, 1952); however, it should be clear that the principles are generalisable to experiments containing larger choice sets.

As Bradley-Terry models specify choice sets to consist of two options, Equation (56) can be simplified to

$$P_{ts} = \frac{\pi_i}{\pi_i + \pi_j}, \quad \text{Equation 57}$$

where j is the option in choice set s which is not i . It is then necessary to identify which pairs of options T_i and T_j are contained within the experiment, which is defined as

$$n_{ij} = \begin{cases} 1 & \text{when the pair } (T_i, T_j) \text{ is in the choice experiment} \\ 0 & \text{when the pair } (T_i, T_j) \text{ is not in the choice experiment.} \end{cases}$$

Additionally, to construct the likelihood function, it is helpful to define the choices of subject α as

$$\omega_{ij} = \begin{cases} 1 & \text{when } T_i \text{ is preferred to } T_j \\ 0 & \text{when } T_i \text{ is not preferred to } T_j. \end{cases}$$

The next step is to define $f_{ij\alpha}(\omega_{ij\alpha}, \boldsymbol{\pi})$ as the probability density function for individual α in choice set (T_i, T_j) where $\boldsymbol{\pi}=(\pi_1, \pi_2, \pi_3, \dots, \pi_t)$. If there are s respondents and the sum of the choices over the α respondents $\sum_{\alpha} \omega_{ij\alpha} = \omega_{ij}$, the likelihood function is given by #

$$L(\boldsymbol{\pi}) = \prod_{i < j} \prod_{\alpha=1}^s f_{ij\alpha}(\omega_{ij\alpha}, \boldsymbol{\pi}) = \prod_{i < j} \left(\frac{\pi_i^{\omega_{ij}} \pi_j^{s\omega_{ij} - \omega_{ij}}}{(\pi_i + \pi_j)^{s\omega_{ij}}} \right). \quad \text{Equation 58}$$

As Street and Burgess (2007) note, this can be simplified further by allowing

$\omega_i = \sum_j \omega_{ij}$ be the total number of times that T_i is chosen, i.e.

$$L(\boldsymbol{\pi}) = \frac{\pi_1^{\omega_1} \pi_2^{\omega_2} \dots \pi_I^{\omega_I}}{\prod_{i < j} (\pi_i + \pi_j)^{sn_{ij}}} \quad \text{Equation 59}$$

It is convenient to take logs to produce the log-likelihood,

$$\ln(L(\boldsymbol{\pi})) = \sum_{i=1}^I \omega_i \ln(\pi_i) - \sum_{i < j} sn_{ij} \ln(\pi_i + \pi_j). \quad \text{Equation 60}$$

To estimate the values of each of the π_i terms, the first derivative of the likelihood function (or equivalently, of the log-likelihood) is set at 0, and solved iteratively. The

$\hat{\pi}_i$ terms are then normalised so that $\prod_i \hat{\pi}_i = 1$. The $\hat{\pi}_i$ terms which meet this restriction are termed the maximum likelihood estimators. Street and Burgess (2007) provide a simple worked example of MLE (pp.62-65). In practice, this process is rarely performed manually; all leading statistical packages (including STATA which is the main software used in this thesis) do this automatically. Greene (2003) notes some of the attractive properties of maximum likelihood estimators (p.473). These four properties are that the estimator is consistent, has asymptotic normality, has asymptotic efficiency, and invariance.

Up to this point, I have discussed the appropriate construction of orthogonal arrays, and the methods for estimating parameters through maximum likelihood estimation. With these building blocks in place, the focus of this section turns to the key issue of how to compare DCEs, which in this thesis pair items from orthogonal arrays with alternative profiles to create a set of choice sets.

Deriving the information (Λ) matrix

The information matrix (termed the Λ matrix from this point) is a matrix of expectations of products of partial derivatives of the log-likelihood function. In the context of pairs (which as noted previously is necessary for the Bradley-Terry model), a design will have sn_{ij} observations from the pair of options (T_i, T_j) , and sN

($= s \sum_{i < j} n_{ij}$) observations in total. The proportion of observations from any choice set

is therefore n_{ij} / N , and will be denoted by λ_{ij} . The elements of the information matrix

are

$$(\Lambda)_{ij} = \sum_{i < j} \lambda_{ij} e_{\pi} \left(\left(\frac{\partial \ln f_{ij}(\omega_{ij}, \pi)}{\partial \pi_i} \right) \left(\frac{\partial \ln f_{ij}(\omega_{ij}, \pi)}{\partial \pi_j} \right) \right) \quad \text{Equation 61}$$

where

$$\lambda_{ij} = \begin{cases} 1/N & \text{when the pair } (T_i, T_j) \text{ is in the choice experiment} \\ 0 & \text{when the pair } (T_i, T_j) \text{ is not in the choice experiment.} \end{cases}$$

The entries in the Λ matrix are therefore

$$(\Lambda)_{ij} = \frac{-\pi_i \pi_j \lambda_{ij}}{(\pi_i + \pi_j)^2} \quad \text{Equation 62}$$

$$(\Lambda)_{ii} = \pi_i \sum_j \frac{\pi_j \lambda_{ij}}{(\pi_i + \pi_j)^2} \quad \text{Equation 63}$$

As these entries include elements of π , it is usual to assume for the purposes of the Λ matrix that all of the terms are equal (and in fact are all one because of the normalising procedure described previously). However, it is also possible to adopt prior information about these values. A simple example of a Λ matrix assuming all of the terms are equal is now presented.

Imagine a situation in which the choice of health insurance is being explored (denoted from this point as **Example 1**). The cost of the insurance has only two levels (high and low, coded as 0 and 1 respectively) and the level of coverage only has two levels (extensive and minimal, coded as 0 and 1 respectively). Thus, there are only four possible programs on offer (00, 01, 10, 11) where the first number reflects the cost of the insurance, and the second the level of coverage). Imagine a very simple choice experiment with only two questions, each with two options. These two questions are (00,11) and (01, 10). For this choice experiment, the Λ matrix can be estimated, with each of the entries representing the number of occurrences of the pairs of profiles in the choice experiment (El Helbawy and Bradley, 1978). It is a 4 x 4 matrix with the rows and columns representing each of the four hypothetical insurance programs. For the off-diagonal positions, a 0 represents that this pair of profiles does not occur in the choice experiment. The Λ matrix for **Example 1** is

$$\Lambda = 1/8 \begin{bmatrix} & 00 & 01 & 10 & 11 \\ 00 & 1 & 0 & 0 & -1 \\ 01 & 0 & 1 & -1 & 0 \\ 10 & 0 & -1 & 1 & 0 \\ 11 & -1 & 0 & 0 & 1 \end{bmatrix}.$$

This is obviously a very simple example, and it is useful to note that Equations (61-3) can be generalised for situations with more than two options per choice set. In this case,

$$\lambda_{ij\dots k} = \begin{cases} 1/N & \text{if } T_i, T_j, \dots, T_k \text{ is in the choice experiment} \\ 0 & \text{otherwise} \end{cases}$$

and the entries in the Λ matrix become

$$(A)_{ij} = -\pi_i \pi_j \sum \frac{\lambda_{ij\dots k}}{(\sum_{l=1}^m \pi_l)^2} \quad \text{Equation 64}$$

$$(A)_{ii} = \pi_i \sum \frac{\lambda_{ij\dots k} \sum_{l=1}^m \pi_l}{(\sum_{l=1}^m \pi_l)^2} \quad \text{Equation 65}$$

This is best explained using an example. Imagine a choice experiment investigating choices surrounding the use of a new vaccine (denoted from this point as **Example 2**). A DCE is constructed with two attributes, with the following levels:

1. Effectiveness of vaccine (90%, 95%, 100%, coded as 0,1,2 respectively)
2. Cost of vaccine (\$20, \$100, \$200, coded as 0,1,2 respectively)

Thus, there are 9 hypothetical vaccines which might be selected, namely 00, 01, 02, 10, 11, 12, 20, 21 and 22 (where the first number reflects the effectiveness of the vaccine, and the second the cost). Imagine a choice experiment with only two questions (choice sets) in it, both presented as triples (i.e. the respondent has only three hypothetical vaccines to select from in each of the two questions). These two questions are (00, 12, 21) and (01, 10, 22). As the π terms in Equation (64) and (65) are assumed to be 1, a -1 indicates that the pair does occur in the experiment. The diagonal components are selected such that the sum of any row or column is equal to 0. The entire matrix is then divided by m^2N where m is the number of options in each choice set (here 3), and N is the number of choice sets (here 2). Thus, the Λ matrix (with row and column labels added for convenience) in this case is:

$$\Lambda = 1/18 \begin{bmatrix} & 00 & 01 & 02 & 10 & 11 & 12 & 20 & 21 & 22 \\ 00 & 2 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 01 & 0 & 2 & 0 & -1 & 0 & 0 & 0 & 0 & -1 \\ 02 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & -1 & 0 & 2 & 0 & 0 & 0 & 0 & -1 \\ 11 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 12 & -1 & 0 & 0 & 0 & 0 & 2 & 0 & -1 & 0 \\ 20 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 21 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 2 & 0 \\ 22 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Note that the experiment is potentially limited in the information it can provide the analyst as three of the profiles do not appear in any of the choice sets (and hence have rows and columns of 0). If we are interested in main effects only, the lack of particular profiles is unproblematic; what matters most is level occurrence rather than profile occurrence. However, as the analyst moves towards estimation of interaction terms, the rows and columns of 0s in the Λ matrix become problematic. The Λ matrix identifies which pairs of profiles occur in the choice experiment. However, it is difficult to determine relatively better designs based on the Λ matrix as it stands. Usually one number summaries of the information matrix are used to compare designs and this idea is developed now.

***B* and *C*-matrices**

The next step in evaluating designs is to generate the B matrix. The B matrix identifies contrasts that can be explored within the given design (for example main effects only, main effects and a subset of two-factor interactions, or main effects and all two-factor interactions). In constructing the B matrix, it is necessary to identify the effects that are of interest to the analyst (unlike the Λ matrix). To be explicit, the Λ matrix is identical whether the experiment focuses on main effects only or on those plus some interaction terms. However, the B and C -matrices are not and reflect the characteristics of the selected choice sets relative to some pre-specified some of effects of interest.

If we begin considering a situation in which only main effects are of interest, the B matrix represents the orthogonal contrasts for each of the main effects of each of the attributes. Thus, there are $l_q - 1$ contrasts for each dimension of the experiment, where

l_q is the number of levels. Each row of the B matrix represents one of these contrasts. The columns of the B -matrix represent the possible combinations of levels within the experiment. For **Example 1**, the B matrix is

$$B = \begin{bmatrix} -1/2 & -1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & -1/2 & 1/2 \end{bmatrix}$$

In **Example 2** introduced above, there will be four rows in the B matrix, representing the linear and quadratic contrasts for the two attributes. This is

$$B = \begin{bmatrix} \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & 0 & 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ 1 & 1 & -\sqrt{2} & -\sqrt{2} & 1 & 1 \\ \frac{3\sqrt{2}}{3\sqrt{2}} & \frac{3\sqrt{2}}{3\sqrt{2}} & \frac{3}{3\sqrt{2}} & \frac{3}{3\sqrt{2}} & \frac{3\sqrt{2}}{3\sqrt{2}} & \frac{3\sqrt{2}}{3\sqrt{2}} \\ \frac{-1}{\sqrt{6}} & 0 & \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{6}} \\ 1 & -\sqrt{2} & 1 & 1 & -\sqrt{2} & 1 \\ \frac{3\sqrt{2}}{3\sqrt{2}} & \frac{3}{3\sqrt{2}} & \frac{3\sqrt{2}}{3\sqrt{2}} & \frac{3\sqrt{2}}{3\sqrt{2}} & \frac{3}{3\sqrt{2}} & \frac{3\sqrt{2}}{3\sqrt{2}} \end{bmatrix}$$

In an ordinary least squares setting, a good estimation procedure for a parameter is often defined as one which produces a minimum variance unbiased estimator. As choice experiments are interested in more than one effect, evaluation analogous to this requires the variance-covariance matrix C^{-1} constructed as $C=BAB'$ (El Helbawy and Bradley, 1978). One attractive characteristic of C^{-1} is that it should be block diagonal for the effects from different attributes. In a 2^k situation, effects can be independently identified only if all off-diagonal positions are zero. If attributes have more than 2 levels, correlations are permissible between (for example) the linear and quadratic terms of an attribute, but not between either of these and any other terms investigated in the matrix. However, it should be noted that this block diagonal property does not always co-occur with the design with the best efficiency (a concept which is discussed below).

The C -matrix for **Example 1** is this,

$$C = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/4 \end{bmatrix},$$

with the corresponding C^{-1} matrix being the inverse, which is

$$C^{-1} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}.$$

As the off-diagonal positions in this case are 0s, the design produces uncorrelated main effects.

The C -matrix for **Example 2** is given below:

$$C = \begin{bmatrix} 1/9 & 0 & 1/18 & 0 \\ 0 & 1/9 & 0 & -1/18 \\ 1/18 & 0 & 1/9 & 0 \\ 0 & -1/18 & 0 & 1/9 \end{bmatrix},$$

with the corresponding C^{-1} matrix being

$$C^{-1} = \begin{bmatrix} 12 & 0 & -6 & 0 \\ 0 & 12 & 0 & 6 \\ -6 & 0 & 12 & 0 \\ 0 & 6 & 0 & 12 \end{bmatrix}.$$

The C^{-1} matrix shows this selection of choice sets does not have uncorrelated main effects. After outlining methods to estimate the statistical efficiency of designs, I will move on to consider how choice sets can be selected in a more systematic way to ensure that the analyst can independently estimate the effects considered most important.

D-efficiency

The C -matrix can be used to evaluate the statistical efficiency of the design of the experiment. D-efficiency, which is the central concept of efficiency considered in this thesis, is defined as

$$\left(\frac{\det(C)}{\det(C_{optimal})} \right)^{1/p}, \quad \text{Equation 66}$$

where p is the number of parameters to be estimated, and C is the information matrix for these p parameters (Street and Burgess, 2007). A significant strength of this approach is that, in the context of main effects (and some limited extensions into estimation of interaction terms) the optimal design need not be known to estimate the

D-efficiency (Louviere, et al., 2003). The reason for this is that Burgess and Street (2007) have provided an upper bound for the determinant of the C matrix for estimating main effects for any choice set size, with any number of attributes and levels:

$$\det(C_{optimal}) = \prod_{q=1}^k \left(\frac{2S_q}{m^2 (l_q - 1) \prod_{i=1, i \neq q}^k l_i} \right)^{l_q - 1}, \quad \text{Equation 67}$$

where

$$S_q = \begin{cases} (m^2 - (l_q x^2 + 2xy + y)) / 2 & 2 \leq l_q \leq m, \\ m(m-1) / 2 & l_q \geq m \end{cases} \quad \text{Equation 68}$$

and positive integers x and y satisfy the equation $m = l_q x + y$ for $0 \leq y < l_q$. Street and Burgess (2005) define S_q to be “the maximum number of differences in the levels of attribute q in each choice set” (p.463)

An upper boundary has only been established for main effects, and when all attributes are binary for the estimation of main effects and two-factor interactions. In the last of these, Street and Burgess (2007) identify the determinant of the optimal design to be:

$$\det(C) = \begin{cases} \left(\frac{(m-1)(k+1)}{mk2^k} \right)^{k+k(k-1)/2} & k \text{ is odd} \\ \left(\frac{(m-1)(k+2)}{m(k+1)2^k} \right)^{k+k(k-1)/2} & k \text{ is even} \end{cases} \quad \text{Equation 69}$$

Publicly available software can be used to calculate the statistical efficiency for specific designs (<http://maths.science.uts.edu.au/math/wiki/SPExptSoftware>). The statistical efficiency for **Examples 1** and **2** can be calculated using this software as 100% and 86.6% respectively.

Alternatives to D-efficiency

It should be noted that there is considerable discussion regarding how best to summarise the C-matrix into a single measure of efficiency for comparison of designs. Street and Burgess (2007) define some of the leading candidates,

“The D-, A-, and E-optimality measures are appropriate to our situation and we now define these...

...A design is D-optimal if it minimizes the generalized variance of the parameter estimates, that is, $\det(C^{-1})$ is as small as possible for the D-optimal designs.

A design is A-optimal if it minimizes the average variance of the parameter estimates, that is, $\text{tr}(C^{-1})$ is as small as possible for the A-optimal designs.

A design is E-optimal if it minimizes the variance of the least well-estimated parameter, that is, the largest eigenvalue of C^{-1} is as small as possible for the E-optimal designs” (p.84)

Note that this definition of D-optimality has reversed the maximisation, in that optimality is defined in terms of minimising the determinant of the inverse; this is of course the same criterion expressed differently. The strength of D-optimality lies in its ability to maximise the predictive value of the design. This differs from A-optimality which minimises the average of the variances of the parameter estimators. The choice of optimisation strategy therefore depends on what the analyst considers most important. In the contexts which will be introduced in the empirical chapters that follow, both predictive value and precise estimators are important. Arguably, the choice between D- and A-optimality is not of great importance. Firstly, Chapter 4 in Street and Burgess (2007) note that, for experiments with binary levels, A- and D-efficiency are the same. Using simulation data, Kessels *et al.* consider both the expected mean square errors of the parameter estimates and of the predicted probabilities (Kessels, et al., 2006). Under both criteria, designs generated under both optimisation strategies have similar results. For this thesis, the default option was to adopt D-optimality as the primary goal of design, although the parameter estimates have to be considered in the context of a slight reduction in precision.

One issue with the use of A-efficiency is that it is dependent on the scale of the parameters. The variance of the parameter estimate is related to the size of the coefficient, which in turn depends on the scale. However, this is not a fatal flaw as comparison between alternative designs is still valid using A-efficiency.

Design strategies

In this design section, I have so far introduced methods for evaluating designs. The next step is to discuss the design strategies that have been widely employed in published DCEs. I will focus on four approaches which I will label the Huber and Zwerina (H&Z) approach (1996), the L^{MA} approach (Louviere, et al., 2000), the use of SAS search algorithms (Kuhfeld, 2010), and the generator-developed approach, extensively investigated by Street and Burgess (2007).

Huber and Zwerina (1996) identify four properties that characterize efficient choice designs. These are level balance, orthogonality, minimal overlap and utility balance. Level balance is satisfied when the levels in a dimension occur with equal frequency. Orthogonality is satisfied when each pair of levels of attributes occurs with equal frequency on the same profile.

Minimal overlap says that there should be as few instances as possible of attributes being at the same level within a particular choice set. This appears plausible; at the extreme, a situation with universal overlap in which two identical options are presented in a choice set, provides no information. However, the concept of minimal overlap becomes more complicated when higher-order effects (interactions) are of interest. In these circumstances, varying one dimension while holding others fixed is necessary despite violating minimal overlap for one attribute. This idea will be explained further in the example which follows in the description of generator-developed designs.

Utility balance asserts that the systematic component of the utility function should value two options within a choice set as equally as possible. Unlike the other three criteria, this makes assumptions about the importance of each level of each dimension. Under an assumption of zero coefficients, it is clear that utility balance is true. Huber and Zwerina (1996) argue that if there is a reasonable expectation of non-zero coefficients, selecting a design with utility balance can reduce the number of respondents needed to achieve a specific error level around the parameter by 10-50%. Kanninen (2002) argues that equal likelihood of selecting options in a choice set is not necessarily optimal; rather it depends on the number of attributes in the experiment. The most unequal optimal utility balance occurs with two attributes. In this case, the optimal probabilities (in terms of maximising the determinant of the Fisher

Information Matrix) are 0.82 / 0.18 (i.e. an 82% chance of selecting Option A). Clearly, this means that non-zero priors have to be assumed (as choices under zero priors are based on identical systematic utility across options within a choice set). Importantly, Street and Burgess (2007) (p.88-90) give an example that shows that satisfying the four criteria of Huber and Zwerina (1996) does not ensure that main effects can be estimated.

L^{MA} designs are derived from an orthogonal main effects plan (OMEPE), and are discussed at length by Louviere *et al.* (2000). The term L^{MA} means there are L levels for each of A attributes for each of the M options. In an OMEPE, each pair of levels of particular dimensions appears with equal frequency allowing independent estimation of the main effects (Dey, 1985). An issue with L^{MA} designs is that they require a large OMEPE since each run is used to define the first and all subsequent options in a choice set. In the simple example presented previously, two dimensions each with three levels were developed into two choice sets with three options. Using L^{MA} it would be necessary to find an OMEPE with at least six three-level dimensions. A $3^6 \times 6$ OMEPE exists; if the six level dimension is dropped, an L^{MA} design for this situation remains and is given below.

Table 15: Example L^{MA} Design

Choice Set Number	Option A	Option B	Option C
1	00	00	00
2	00	11	22
3	01	02	21
4	01	20	12
5	02	12	10
6	02	21	01
7	10	02	12
8	10	20	21
9	11	11	11
10	11	22	00
11	12	01	20
12	12	10	02
13	20	12	01
14	20	21	10
15	21	01	02
16	21	10	20
17	22	00	11
18	22	22	22

Note that this particular L^{MA} design has a number of troubling characteristics. Some choice sets present the same option in all three options (numbers 1, 9 and 18). Some choice sets are identical (2, 10 and 17). This would suggest that this approach requires the use of a carefully selected OMEP, or the use of a labelled experiment (which effectively introduces an additional attribute).

L^{MA} designs are often appropriate for use with labelled experiments and overcome this repetition of profiles without needing to find a different OMEP or manipulating an existing one. In a labelled design, the option lettering (i.e. the A, B, C terms above) refer instead to a specific product such as bus, car, train in a transport example, or Woolworths, Coles, IGA in a supermarket example.

The idea of shifted designs central to Street and Burgess (S&B) designs was introduced by Bunch *et al.* (1996). In this, a set of initial options is chosen for each of the choice sets in the experiment, and the subsequent option or options are defined by modular arithmetic to “*shift each combination of initial attribute levels by adding a constant that depends on the number of levels*”. Thus, subsequent options in each choice are defined by these constant shifts, and the set of constant shifts required for all attributes is called a generator.

Generally, the initial options in each choice set are obtained from an orthogonal main effects plan (OMEPE). Imagine a simple choice experiment with three dimensions, each with only two levels coded as 0 and 1. This may represent the choice of health insurance policy, where the three attributes relate to price, coverage for dental care and for obstetrics respectively. OMEPEs with 4 runs exist for this 2^3 design, one being (0, 0, 0), (0, 1, 1), (1, 0, 1) and (1, 1, 0).

To generate the second (or subsequent) option in each choice experiment, the generator is selected to allow for investigation of main effects. If we want to know the importance of changing levels within a dimension, it is valuable to have choice sets in which the dimension occurs both at level zero and at level one. Therefore, an obvious choice for the generator is (1, 1, 1). Thus, the choice experiment would have four choice sets, as given below.

Table 16: Example Main-Effects Only Choice Experiment

Option A	Option B
(0, 0, 0)	(1, 1, 1)
(0, 1, 1)	(1, 0, 0)
(1, 0, 1)	(0, 1, 0)
(1, 1, 0)	(0, 0, 1)

The Λ matrix and the C matrix for the estimation of main effects for the simple design in Table 16 are provided below:

$$\Lambda = 1/16 \begin{bmatrix} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ 000 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 001 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 010 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 011 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 100 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 101 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 110 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 111 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$C = \begin{bmatrix} \frac{1}{8} & 0 & 0 \\ 0 & \frac{1}{8} & 0 \\ 0 & 0 & \frac{1}{8} \end{bmatrix}$$

Given that the example was based on estimation of main-effects alone, the two matrices show the design approach was successful. In the Λ matrix, the data show each combination of dimensions and levels is represented in the choice experiment. As noted previously, the importance of missing combinations is dependent on the effects of interest. In the C matrix, the off-diagonal positions are all zero; therefore all effects can be estimated independently. The statistical efficiency of the design is 100%.

A limitation of the approach taken by Bunch *et al.* (1996) was that it focused on estimation of main effects only. However, Street and Burgess (2007) extended the idea of using generators by showing that choosing generators with certain structural properties can allow for estimation of specific interaction terms. The choice of these shift generators is determined by the interaction effects the analyst wishes to investigate.

To allow estimation of interactions, the choice of generators becomes more difficult. Suppose we use the design presented in Table 16 to investigate two-factor interactions in addition to main effects. In this case, the C matrix becomes

$$C = \begin{bmatrix} \frac{1}{8} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus, the interaction terms cannot be estimated, and the statistical efficiency of the design for this context is 0%. To allow the experiment to estimate all terms, a different approach to selecting both the starting design and generators is required. Using an OMEP as a starting design when interactions are of interest is inappropriate; rather, a starting design of resolution 5, equivalently strength 4 is required. Regarding the selection of generators, for an interaction between two attributes to be estimable, the generators must have different levels in the two positions (for the binary case, this means a 0 and a 1 in the two positions). While the generators have to be selected to allow estimation of all relevant effects, additional generators will increase the size of

the experiment, and hence the required number of respondents (or number of responses per respondent). Street and Burgess (2007) provide Lemma 4.2.4. which identifies that, for a design with only binary attributes,

*“If $2^m \leq k < 2^{m+1}$, then there is an estimable set with $m+1$ generators”
(p.129)*

Thus, in this simple experiment in which $k = 3$, we know that two generators is adequate to explore all main effects and two-factor interactions. In selecting the specific generators, Street and Burgess (2007) state that,

“For the estimation of main effects and two-factor interactions in the complete factorial, generators of weight $(k+1)/2$ have been shown to be optimal for odd k . For even k , generators of weight $k/2$ and $k/2 + 1$ have been shown to be optimal.” (p.129)

This result is intuitive, and stems from Theorem 4.1.3 (Street and Burgess, 2007). As previously stated, for an interaction term between two attributes to be estimable, the generator must have a zero and a one in the corresponding positions. Therefore, the number of interactions that can be investigated using a generator consisting of zeros and ones is $g_0(k - g_0)$ where g_0 is the number of positions in the generator that are equal to zero. This is maximised when $g_0 = k/2$; therefore a weight of approximately $k/2$ is optimal for estimating interaction terms. However, as seen in the previous example, the information provided for the estimation of main effects increases with the number of ones in the generator.

Returning to the starting design, as noted previously, Sloane provides a library of orthogonal arrays including many with strength three or greater (<http://www2.research.att.com/~njas/oadir/>). In the health insurance example introduced previously, the most appropriate array has four two-level attributes, and has eight rows (i.e. half of the full factorial). However, since the example does not require the final two-level attribute, this is dropped leaving the 2^3 full factorial (i.e. 000, 001, 010, 011, 100, 101, 110, 111).

A potential set of generators to estimate interactions might be (101, 110). The reason why this is a potentially good set of generators to be estimated can be shown by

considering which generators provide information on which main effect and interaction terms.

Table 17: Selecting generators to estimate main effects and interactions

Generator	A	B	C	AB	AC	BC
101	√		√	√		√
110	√	√			√	√

We see that each effect of interest has information stemming from choice sets derived by at least one of the generators. The choice of generators is partially based on the effects which are most of interest; for example, the main effect on A is well explored, while those of B and C are less so. Because of duplication (and subsequent removal) of choice sets, this design contains only 8 choice sets. The Λ and C-matrices for main effects and two factor-interactions are presented below.

$$\Lambda = 1/32 \begin{bmatrix} & 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ 000 & 2 & 0 & 0 & 0 & 0 & -1 & -1 & 0 \\ 001 & 0 & 2 & 0 & 0 & -1 & 0 & 0 & -1 \\ 010 & 0 & 0 & 2 & 0 & -1 & 0 & 0 & -1 \\ 011 & 0 & 0 & 0 & 2 & 0 & -1 & -1 & 0 \\ 100 & 0 & -1 & -1 & 0 & 2 & 0 & 0 & 0 \\ 101 & -1 & 0 & 0 & -1 & 0 & 2 & 0 & 0 \\ 110 & -1 & 0 & 0 & -1 & 0 & 0 & 2 & 0 \\ 111 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

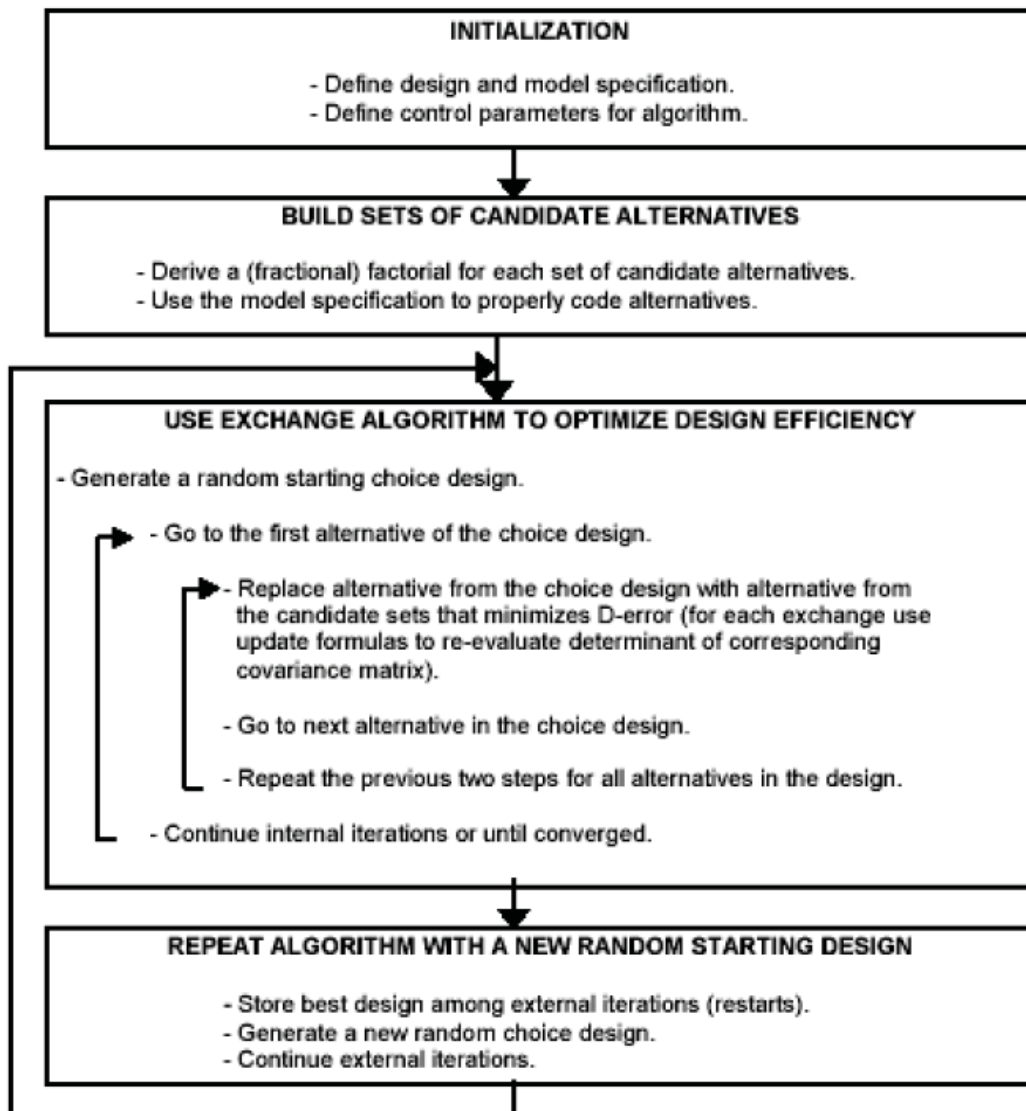
$$C = \begin{bmatrix} \frac{1}{8} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{16} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{16} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{16} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{16} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{8} \end{bmatrix}$$

The matrices suggest this is an appropriate design for this example. The off-diagonal positions are all zero. The efficiency of the design is 94.49%. Note that Table 4.12 in Street and Burgess identifies that 100% efficiency can be attained using (011, 101, 110) as generators.

SAS Algorithms

An alternative strategy to designing experiments which will be used in the next chapter (the SF-6D DCE) is the use of the SAS search algorithms (Kuhfeld, 2010). The basic approach taken by these algorithms is to generate an OMEP and then, using a user-specified number of treatment combinations, searches for choice sets using a search algorithm. A section of Kuhfeld's online book, authored by Zwerina, Huber and Kuhfeld (2010), details this approach. The approach begins with a list of potential alternatives, of which a random selection is chosen. The approach then finds the best exchange for the first alternative in this random selection of choice sets (in terms of the exchange which maximises D -efficiency). This is repeated for the next alternative, and so on until the algorithm has sequentially found the best exchanges for all alternatives. This entire process is then repeated until no substantial improvement in D -efficiency is possible. This process can be repeated multiple times with different starting designs to avoid poor local optima. This process is summarised in Figure 14, which is a reproduction of Figure 1 of Zwerina *et al.* (2010).

Figure 14: Flowchart of algorithm for constructing efficient choice designs



Street, Burgess and Louviere (2005) and Street and Burgess (Chapter 8) (2007) have investigated the appropriateness of the choice sets this method can provide. They conclude that the design generated using this approach has good efficiency; however, it does not ensure uncorrelated estimates of the effects of interest.

Some other areas of interest

While generator-developed designs (which are used in this thesis) are only guaranteed to be optimal in terms of D -efficiency when the coefficients are zero, an issue remains about how best to make assumptions regarding coefficients which would potentially allow a better design to be constructed. One option is to update the prior values sequentially, by adjusting the design as data are generated (Kanninen, 2002). There are a range of other studies that have investigated the benefits of

designing experiments based on non-zero priors (Arora and Huber, 2001; Ferrini and Scarpa, 2007). As the work undertaken in this thesis will generally focus on welfare measures, an important result is that of Carlsson and Martinsson (2003), who argue that these welfare measures are not significantly affected by using incorrect priors within a D-efficient design. Nevertheless, this is beyond the scope of this thesis and I will focus on the efficiency of designs based on zero priors.

One additional similar issue is that some recent studies have investigated design of experiments aimed at investigating not just mean response, but also respondent heterogeneity. Drawing on Bayesian principles, a number of studies have attempted to design choice experiments for the mixed logit approach (Regier, et al., 2009; Sándor and Wedel, 2005; Yu, et al., 2009), all of which identify that considering the heterogeneity of responses when designing choice experiments improves the efficiency of the design.

One final question that I would like to note is the issue of how many alternatives should be presented in each choice set. In efficiency terms, Burgess and Street (2006) identified that choice sets with two alternatives are unlikely to maximise the determinant of the C-matrix. However, as the authors note,

“(P)ractitioners will need to decide how to trade off gains in statistical efficiency with potential losses in respondent efficiency.”(p.515)

This is an important point and relates back to the cognitive burden of the task discussed previously. In a revealed preference setting, there is a broad literature concerning the difficulties posed by increasing choice (Boatwright and Nunes, 2001; Iyengar and Lepper, 2000). Arguably, the appropriate comparison between choice sets of different sizes would be to assume the sample size in an experiment to be inversely proportional to the size of each choice set, and then to determine if the smaller sample associated with the more statistically efficient design outperforms the larger sample answering the less statistically efficient design in terms of precision and bias in the point estimates. This is beyond the scope of the work presented here, and the choice experiments presented in the following empirical chapters consider only pairs. However, this does not imply that these are superior to larger choice sets.

Chapter summary

Regarding design of experiments, I concluded that the use of shift generators or SAS algorithms was appropriate, particularly given the difficulties in extending other leading design strategies (L^{MA} , Huber and Zwerina) to deal with interaction terms. For the investigation of heterogeneity, it was concluded that it was valuable to extend beyond a simple random-effects probit or logit, but the methods for using these heterogeneity results in health policy was uncertain. A specification of the utility function was proposed which allows for the inter-related nature of the dimensions of the choice experiments that now follow in the empirical chapters of this thesis.

Chapter 5: Using a Discrete Choice Experiment to Value Health Profiles in the SF-6D

Chapter summary

This chapter presents an empirical study designed to elicit weights for the SF-6D using a DCE. The data were collected as part of a larger project (NHMRC Project Grant 403303). The methods of data collection, and rationale for various decisions made in that process, are outlined to provide background to the analysis section. The approach to data analysis was developed and performed as part of this thesis.

This chapter first briefly reintroduces Random Utility Theory (RUT) as the foundation of discrete choice experiments, and describes some existing studies which have attempted to use a DCE approach to value generic multi-attribute utility instruments. Then, it describes the data that are used in this chapter, including the specific employment of design strategies, the approach to sampling, the base case random-effect probit results, and the resulting QALY weights for the SF-6D. These are contrasted with existing weights for the SF-6D from the United Kingdom. An additional comparison is made between these Australian DCE-derived SF-6D weights with a set of Australian DCE-derived EQ-5D weights, with the intent to explore whether using a common valuation approach reduces the divergence between EQ-5D and SF-5D weights introduced and discussed in Chapter 2. Following this, the chapter uses the SF-6D DCE data to explore whether people differ in their responses based on observable characteristics. For instance, is the valuation of a specific health state dependent on the gender of the respondent? Finally, following the approach described in detail in Chapter 3, using newly developed STATA code (Gu, et al., 2011), I consider different approaches to modelling heterogeneity ranging from a simple conditional logit through to the generalised multinomial logit, which accounts for both scale and preference heterogeneity (Fiebig, et al., 2010).

Introduction – Using ordinal data to value health states

As discussed in Chapter 3, the use of discrete choice experiments relies on the concept of random utility. In this, the utility of an alternative i in a choice set C_n to an individual n is given by

$$U_{in} = V_{in}(X_{in}, \beta) + \varepsilon_{in} \quad \text{Equation 70}$$

Thus, U_{in} consists of a systematic component $V_{in}()$ and an error term ε_{in} . If there are J items in C_n , the choice is defined by

$$y_{in} = f(U_{in}) = \begin{cases} 1 & \text{if } U_{in} = \max_j \{U_{ij}\} \cdot \forall j \neq i \in C_n \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation 71}$$

Therefore, alternative i is chosen if and only if

$$(V_{in} + \varepsilon_{in}) > (V_{jn} + \varepsilon_{jn}) \cdot \forall j \neq i \in C_n \quad \text{Equation 72}$$

The terms in these equations are not observed; the only observed term is the preference between the two composite terms. Therefore, analysis is reliant on inferring the terms from that relative preference (which is observed). The model is therefore probabilistic as the error term is unbounded and therefore able to reverse any preference implied by the systematic component of the utility function (Marschak, 1960; McFadden, 1981). The methods used in these experiments are consistent with Random Utility Theory (RUT) in economics and psychology (which was discussed in the introductory chapter) and with a Lancasterian approach to consumer theory (Lancaster, 1966; Thurstone, 1927a) in which the utility of a good consists of the utility associated with its various characteristics. These two issues, RUT and a Lancasterian approach to consumer theory, are distinct but both contribute to the DCE field in terms of how we describe individual utility functions, and how these are then used to predict choices.

Applications of DCEs to value health profiles

Applications in health have been relatively recent, with papers dating from 1990 (Propper, 1990). In the past ten years there has been a rapid growth in the use of this approach in health economics, and there are now many studies using DCEs in a range of applications (de Bekker-Grob, et al., 2012). One strength of DCEs is that, because they are based on the random utility model, they provide a robust theoretical and statistical framework to test the form of the utility function. In particular, the discrete choice experiment approach provides greater flexibility in terms of estimation of flexible functional forms, taking account of interaction effects. Under a DCE approach, the analyst can ensure that interaction effects can be explicitly allowed for in the design, although this can have implications for the sample size required by the analyst.

An important emerging area of application of DCEs in health economics is in the estimation of preferences for health outcomes, and particularly, investigation of the trade-off between quality of life and extra life expectancy, and the nature of the utility function defined over quality of life, survival and other outcomes (Viney, et al., 2002). However, there are relatively few examples in which DCEs have been used to measure the trade-off between quality and quantity of life, or between different dimensions of quality of life. Some examples occur in disease-specific contexts (Gan, et al., 2004; Osman, et al., 2001; Sculpher, et al., 2004). However, few of these studies provide measures of the trade-off between quality of life and survival, a requirement for the construction of the QALY. Hakim and Pathak (1999) use a discrete choice experiment to investigate preferences for EQ-5D health states, but their experiment does not include a time or cost attribute which could be used as a numeraire, thus their results provide information about strength of preference and interaction for EQ-5D dimensions, but not a cardinal measure of preference that could be used to generate a QALY weight (Flynn, et al., 2008). McCabe *et al.* (2006), and Salomon *et al.* (2003) do likewise, using data collected as part of the original EQ-5D and SF-6D valuation studies (Brazier, et al., 2002; Dolan, 1997), although they also do not include a numeraire. These data were collected without consideration of the design of experiments, which can impact on the accuracy and precision of estimated effects (Street and Burgess, 2007).

A Dutch study, already discussed in Chapter 3, attempted to derive utility weights for the EQ-5D using a discrete choice experiment (Stolk, et al., 2010). While the use of DCEs is a step forward on the use of conventional ranking data, the paper by Stolk *et al.* is weakened as it does not include life expectancy in the design of the experiment. In Chapter 3, it was argued that this is an important omission as QALY weights are driven by the willingness of respondents to sacrifice length of life for quality of life, something which is not investigated in an experiment in which only the latter of these is included.

A recent Canadian study (currently available as a working paper only) derived utility weights for the EQ-5D using an approach very similar to that outlined in Chapter 3 (Bansback, et al., 2012). This introduced the specification of the utility function which imposes the zero-condition on the data, and included time as a numeraire. The work presented in this chapter builds on the work of Bansback *et al.* in a number of ways:

1. It uses a regression technique that better accounts for the panel nature of the data
2. It explores approaches to modelling response heterogeneity, including their impact on mean responses
3. It uses a SAS-generated experimental design to ensure unbiased estimates of parameters of interest
4. It produces results applicable to Australian cost-utility analysis

This chapter attempts to remedy some of the limitations that have been identified in the existing literature base. Specifically, an experiment was conducted in which duration was included as variable, different approaches to the modelling of heterogeneity were investigated, and the experiment was designed to ensure both unbiased and precise point estimates for coefficients. This was done in the context of the SF-6D, rather than using the EQ-5D as does much of the existing literature. This does not necessarily reflect the relative merit of the two instruments as I have previously argued that the better descriptive ability of the SF-6D is counterbalanced by the greater difficulty in estimating utility weights for each of the possible states.

The SF-6D

The SF-6D has been described in depth in Chapter 2. To summarise, it is a multi-attribute utility instrument which can be derived from the SF-36 (Ware and Sherbourne, 1992) or the SF-12 (Ware, et al., 1996) for use in economic evaluations. Rather than the 36 items contained within the original instrument, the SF-6D has only six, making administration considerably less onerous. The SF-6D is reproduced in Table 5.

Table 18: The SF-6D

Dimension	Level	
Physical Functioning	1	Your health does not limit you in <i>vigorous activities</i>
	2	Your health limits you a little in <i>vigorous activities</i>
	3	Your health limits you a little in <i>moderate activities</i>
	4	Your health limits you a lot in <i>moderate activities</i>
	5	Your health limits you a little in <i>bathing and dressing</i>
	6	Your health limits you a lot in <i>bathing and dressing</i>
Role Limitation	1	You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems

	2	You are limited in the kind of work or other activities as a result of your physical health
	3	You accomplish less than you would like as a result of emotional problems
	4	You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems
Social Functioning	1	Your health limits your social activities <i>none of the time</i>
	2	Your health limits your social activities <i>a little of the time</i>
	3	Your health limits your social activities <i>some of the time</i>
	4	Your health limits your social activities <i>most of the time</i>
	5	Your health limits your social activities <i>all of the time</i>
Pain	1	You have <i>no</i> pain
	2	You have pain but it does not interfere with your normal work (both outside the home and housework)
	3	You have pain that interferes with your normal work (both outside the home and housework) <i>a little bit</i>
	4	You have pain that interferes with your normal work (both outside the home and housework) <i>moderately</i>
	5	You have pain that interferes with your normal work (both outside the home and housework) <i>quite a bit</i>
	6	You have pain that interferes with your normal work (both outside the home and housework) <i>extremely</i>
Mental Health	1	You feel tense or downhearted and low <i>none of the time</i>
	2	You feel tense or downhearted and low <i>a little of the time</i>
	3	You feel tense or downhearted and low <i>some of the time</i>
	4	You feel tense or downhearted and low <i>most of the time</i>
	5	You feel tense or downhearted and low <i>all of the time</i>
Vitality	1	You have a lot of energy <i>all of the time</i>
	2	You have a lot of energy <i>most of the time</i>
	3	You have a lot of energy <i>some of the time</i>
	4	You have a lot of energy <i>a little of the time</i>
	5	You have a lot of energy <i>none of the time</i>

For the requirements of economic evaluation and the construction of the QALY, the 18,000 possible health states must be placed on a scale with full health equal to one, and health states equivalent to immediate death placed at zero. For the SF-6D, this has been done using the Standard Gamble in a variety of countries to reflect local attitudes to aspects of ill health (Brazier, et al., 2002; Brazier, et al., 2009; Gonçalves Campolina, et al., 2009).

The method used by these studies to value these 18,000 health states has been discussed and critiqued in Chapter 2. Briefly, a selection of 249 of the health states was valued using a Standard Gamble instrument, with the remaining 17,751 being valued using regression techniques. Issues with the use of Standard Gamble, with the selection of the 249 health states, and with the techniques used to value the remaining states have been raised each of which can either artificially bias scores or limit the types of utility functions that can be identified.

The discrete choice experiment methods described in Chapters 3 and 4 are clearly of potential application here. An appropriate design for the experiment can be chosen to allow for a wide range of possible utility functions. Providing ordinal data (rather than a series of responses to a cognitively challenging task such as the Standard Gamble) is likely to allow relatively easy completion of the survey. Designing the experiment to allow estimation of all coefficients of interest precisely and without bias is likely to be of considerable value also. The following sections describe the slight amendment to the SF-6D for the purpose of this work, the construction of the choice experiment, the data collection process and the approach taken to analysis.

The vitality dimension

Before describing the project that collected the data that are analysed in this chapter, it is important to note a slight difference between the conventional SF-6D and the modified SF-6D used in this analysis. Under the modified SF-6D, the wording within the Vitality dimension of the SF-6D was amended to make it easier to understand for survey respondents. The original wording results from the layout of the SF-36, and does not represent how the idea would naturally be expressed. The original and replacement wording are shown in Table 19.

Table 19: The Vitality Dimension

Level	Original Wording	Replacement Wording
1	You have a lot of energy all of the time	You always have a lot of energy
2	You have a lot of energy most of the time	You usually have a lot of energy
3	You have a lot of energy some of the time	You sometimes have a lot of energy
4	You have a lot of energy a little of the time	You rarely have a lot of energy
5	You have a lot of energy none of the time	You never have a lot of energy

The impact of this change was not tested; however it was likely to be small.

Implausibility of health states

Implausibility of health states is important as respondents may have difficulty responding to health states that combine levels of dimensions that seem unlikely. The decision to restrict the experiment to a subset of health states from the instrument is a difficult decision as, while there is a good reason for not presenting implausible health states, this has implications for the statistical efficiency of the design of the experiment. The states defined as implausible in this thesis combined Role Limitations Level 1 (i.e. “You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems”) with Pain Level 6 (i.e. “You have pain that interferes with your normal work (both outside the home and housework) extremely”). There were a number of other pairs which represented unlikely combinations of states. However, these were not excluded because of this balance between presenting only the most plausible comparisons and the statistical efficiency of any experiment.

Design and presentation of experiment


Using this conservatively constrained set of plausible states, a Discrete Choice Experiment was designed, allowing estimation of all main effects and both linear and quadratic interactions between each attribute and life expectancy. Note that, as the main effect of a dimension at a particular level is estimated using the interaction of life expectancy and that level (rather than the main effect), this limits the analysis to main effects. To estimate two-factor interactions between levels of the SF-6D, the

experiment would have to be designed to capture three-factor interactions involving life expectancy.

To represent a range of life expectancies broad enough to capture non-linearity of preference with regard to time, but not to be unrealistic for older respondents, a range of life expectancies was specified between 1 and 20 years (with the levels in the experiment being 1, 2, 4, 8, 12, 16 and 20 years). The range was selected to be similar to the range considered in a TTO, and to be realistic for most respondents. The choice sets were designed with two health profiles for the respondent to select between, but presented as triples in which the two combinations of health states and life expectancy were placed alongside immediate death. The task for the respondent was to identify which of the three options was considered the best, and which the worst, thus providing a complete ranking within each choice set. An example choice set is given in Figure 15.

Figure 15: An Example Choice Set

If you had to choose between the following states:

			
	State 1	State 2	Immediate death
Physical Functioning	Your health limits you a little in bathing and dressing	Your health does not limit you in vigorous activities	
Role Limitation	You are limited in the kind of work or other activities as a result of your physical health	You accomplish less than you would like as a result of emotional problems	
Social Functioning	Your health limits your social activities all of the time	Your health limits your social activities none of the time	
Pain	You have pain that interferes with your normal work (both outside the home and housework) a little bit	You have no pain	
Mental Health	You feel tense or downhearted and low most of the time	You feel tense or downhearted and low most of the time	
Vitality	You always have a lot of energy	You sometimes have a lot of energy	
Duration	12 years, followed by death	8 years, followed by death	
Which option is the best?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Which option is the worst?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

As noted in Chapter 3, the Immediate Death option was included with the intention of using it to anchor health states on to the 0-1 scale required for economic evaluation. However, for reasons described in Chapter 3, this analysis approach was not taken

and, rather, the relative rank of the two non-death states was used to run the regression, and an alternative approach to anchoring was established which did not require the ranking of health states relative to Immediate Death.

The full factorial for the 6 dimensions of the SF-6D and the 7 life expectancies contained $6^2 \times 4 \times 5^3 \times 7 = 126,000$ health profiles (or 120,750 once the implausible combinations were excluded). Therefore, a design based on an orthogonal fractional factorial was developed. The final design consisted of 180 choice sets. This was generated using the SAS algorithms presented in Kuhfeld (2010), and described in detail by Zwerina *et al.* (2010). These were discussed in Chapter 4, and kindly generated by Dr. Leonie Burgess. This design is given in Appendix 3. Relative to the best identified design using all SF-6D states (i.e. that which maximises the determinant of the *C*-matrix), the design was 99.44% efficient (Street and Burgess, 2007), the small reduction from 100% efficiency reflecting the conservative approach to exclusion of states. There is evidence that up to 16 choice sets is both acceptable to respondents and does not significantly affect responses (Coast, et al., 2006; Hall, et al., 2006a). Therefore, the 180 choice sets were divided into 12 blocks of 15 choice sets, to which the respondents were randomly assigned (but controlled to ensure equal numbers of total respondents per block).

Data and sample recruitment

An online panel of respondents was used for the survey. Respondents were recruited through a large Australia-wide panel provider (Pure Profile Pty). These respondents were each paid a small sum (approximately \$15) to complete the survey. To allow comparability with the Australian population, the panel consists of respondents in line with Australian norms for age and gender. Each respondent used a web link to access the survey, so were able to self-complete at their convenience. To aid the respondent, a thorough description of the task was provided at the beginning of the survey and a help button was available throughout the task. This provided information on how to respond. Each respondent was familiarised with the SF-6D by being asked to describe their own health using the tool (the results from this form part of the data used in the comparison of instruments presented in Chapter 2). Respondents then completed the task for the 15 choice sets. Following this, they answered a series of personal questions regarding gross household income, ethnicity, country of birth, number of

dependents, level of education, age and gender (screen shots of the experiments are provided in Appendix 4). Finally, they were asked how difficult the task was, selecting one of five levels of difficulty ranging from very difficult to very easy. They were also given the opportunity to provide a free-text response outlining their impression of the survey.

Analysis

The purpose of data analysis was two-fold. The first output is a set of utility weights for the SF-6D which can be used in economic evaluation. The second component of the chapter is an exploration of the heterogeneity of responses. In general, the former set of results is produced using population means; extreme preference patterns are only important in so far as they impact on the mean more than less extreme patterns. However, both issues are important. If population weights are constructed using population means, the study of heterogeneity gives an indication of how appropriate the weights are for modelling individual preferences, and also for identifying the degree to which society is in agreement about the importance of different aspects of health.

The first step in the analysis was that, as a data investigation tool, marginal frequencies were estimated. Thus, the probability of picking a health profile given that a particular dimension was at a particular level was identified. The marginal frequency results are likely to be strongly associated with the econometric modelling results; thus, they are a valuable corroborative tool.

The intention had been to anchor the QALY weights using death (which was an option in each of the choice sets). With three options in the choice set, asking the respondent for best and worst profiles produces a complete ranking, which can be exploded out to produce three pairwise preferences. This explosion was proposed by Chapman and Staelin (1982) as a means of gaining additional information. However, the use of the death state as an anchor was ultimately rejected on the grounds that death is considered very differently to health states, and respondents may not consider it within a random utility theory framework. Specifically, as discussed in Chapter 3, respondents may have an *a priori* belief that death cannot be worse than any health state and therefore will never pick it (Flynn, et al., 2008). The framework of the choice experiment did not allow this lexicographic preference to be identified; it was

not possible to identify those respondents who would never prefer death to a health state distinct from those who simply did not see a state that was worse than death. Therefore, an alternative analysis method was adopted in which only the pairwise preference between the two non-death options was considered. The utility of death was simply the utility when time was set at zero (thus, the systematic component of the utility function is zero).

With one exception relating to the number of parameters considered to be random (which is discussed below in the section titled *Limitations in Specifying Random Parameters*), the analysis followed the strands described in Chapter 3. Thus, in STATA, the base case model employed random-effects probit (and random-effects logit as a check for consistency of conclusions between logit and probit models). Then, twelve models were estimated to investigate response heterogeneity. The base case model A plus the heterogeneity investigation models A1-A6 used a QALY-type model in which the utility of alternative j in scenario s for individual i was

$$U_{isj} = \alpha LIF E_{isj} + \beta X'_{isj} LIF E_{isj} + v_i + \varepsilon_{isj} \quad \text{Equation 73}$$

Thus, the sole main effect was on the life expectancy (*LIFE*), and the other characteristics enter the utility function as interactions with the *LIFE* attribute. As noted in Chapter 3, this is an important amendment to a simple additive utility function as it imposes the zero-condition in which all options where health gain was zero were equally likely to be selected within a choice set irrespective of the characteristics of the hypothetical person ‘receiving’ it. Additionally, as *LIFE* only enters as a linear term, it assumes linearity of utility with respect to time.

Within this framework, four models were estimated. Initially, **Model A** X'_{isj} included only the 25 main effects, these being the movement from level 1 to any other level in each of the 6 dimensions of the SF-6D.

The SF-6D is not strictly monotonic in the way that other instruments such as the EQ-5D are. That is, there are some dimensions within which health is not necessarily worse at higher levels. For example, it is unclear if Role Limitation level 2 (“You are limited in the kind of work or other activities as a result of your physical health”) is preferable to Role Limitation level 3 (“You accomplish less than you would like as a result of emotional problems”). While there are instances of uncertain monotonicity in

the SF-6D, there are clearly some pairs of levels within dimensions which are intended to be monotonic. Therefore, if this monotonicity was violated, the violating pairs would be combined, and the model re-estimated as **Model B**. As an alternative to these two models, an additional *MOST* variable was added, which is a dummy equal to one if and only if the health profile has a dimension at the worst level. The reason for running this model is to provide a set of results comparable to those reported by Brazier *et al.* (2002). The results from this are presented as **Model C**.

A series of different parameterisations, particularly in reference to the error term were considered. The models A1-A6, described in detail in Chapter 3, were paired with the simplest main effects model, giving these six models:

- A1: The Conditional Logit
- A2: The Scale Multinomial Logit
- A3: The Mixed Logit (Uncorrelated Coefficients)
- A4: The Generalised Multinomial Logit (Uncorrelated Coefficients)
- A5: The Mixed Logit (Correlated Coefficients)
- A6: The Generalised Multinomial Logit (Correlated Coefficients)

The reason for ordering the models in this way was that the number of parameters increased monotonically from A1 to A6. The reason why these are only run using Model A is that the purpose of modelling heterogeneity in this context is to identify if it improves model fit, which can be done using any of Models A-C. Model A was selected as it was the simplest main effects model so was most likely to achieve convergence. It would be possible to run comparable Models B1-B6 or C1-C6, but achieving convergence is difficult and there is no reason for thinking that patterns of heterogeneity would differ across what are similar models.

Non-linearity in the utility function (models D and D1-D6)

In Models A, B, C, and A1-A6, this thesis considers a utility function which is linear with respect to gain in life expectancy (and when coupled with the zero condition, this forms the QALY model). This is a strong assumption, and requires testing. Utility Function D extends on Utility Function A by relaxing the assumption of linearity of utility with respect to time. Thus, **Model D** is estimated as

$$U_{isj} = \alpha LIFE_{isj} + \rho LIFE_{isj}^2 + \beta X'_{isj} LIFE_{isj} + \phi X'_{isj} LIFE_{isj}^2 + \varepsilon_{isj} \quad \text{Equation 74}$$

74

Thus, the linearity of utility with respect to time is relaxed, as reflected in the $\rho LIFE_{isj}^2$ term in Equation (74). In addition, it relaxes the assumption that the change in total utility associated with it being received by a different group of hypothetical respondents is independent of the total gain (the $\phi X'_{isj} LIFE_{isj}^2$ term).

Models D and D1-D6 are therefore replications of models A and A1-A6 respectively, but adopting this more relaxed non-linear utility function. A summary of the various models being run is provided in Table 20.

Table 20: Models Run in Chapter 5

		Utility Function 1			Utility Function 2
		A	B	C	D
	RE Probit / Logit				
Heterogeneity modelling	Conditional logit	A1			D1
	Scale MNL	A2			D2
	Mixed logit	A3			D3
	G-MNL	A4			D4
	Mixed logit (correlated)	A5			D5
	G-MNL (correlated)	A6			D6

Note that models with this allowance for non-linearity require a large number of additional parameters, and that the number of additional parameters increases substantially as the analysis moves away from the more restrictive models. As described in Chapter 3, model evaluation will primarily be undertaken using the Akaike and Bayesian information criteria (AIC and BIC) (Akaike, 1974; Schwarz, 1978). These consider both the model fit and also the parsimony of the model (by accounting for the number of parameters in the model). The BIC focuses relatively more on parsimony; therefore, disagreement concerning preferred specification (defined by minimising the coefficient) is possible between the two. As noted in Chapter 3, there is an issue with the use of BIC in panel data with multiple observations per person. The question is whether the n term refers to the number of choice sets or the number of respondents. Therefore, both BIC estimates are

calculated, and any disagreement in ranking of models between them are discussed further.

Limitations in specifying random parameters

Since the analyses were conducted in STATA, a technical limitation of the software in this context needs to be noted. For models with random parameters, STATA limits the user to 20 of these. Under Utility Function A (Equation (73)), there are 26 possible random parameters (5 for physical functioning and pain, 4 for social functioning, mental health and vitality, and 3 for role limitation, plus 1 for duration). Initial testing of the importance of making each of these random (by looking at the statistical significance of the standard deviation term) suggested that the coefficients on poorer levels of health were better described as being drawn from a distribution. Therefore, the initial approach was that, for models A3-A6, the random parameters were assumed to be duration and the poorest level within each SF-6D dimension.

Therefore, there were 7 random parameters (duration, physical functioning and pain level 6, social functioning, mental health and vitality level 5, and role limitation level 4). For models D3-D6, the same 7 parameters were assumed to be random, and the quadratic term on duration was added (making 8 random parameters). Some testing of including extra parameters as random was undertaken; in general, the likelihood and speed of achieving convergence reduced considerably.

Rescaling scores for economic evaluation

For the purpose of economic evaluation, it was necessary that the scores attributed to each of the generic quality of life states be on a scale with death at 0 and full health at 1. The initial intention in the empirical chapters was to use the relative preference for the health state and an Immediate Death health state to do this. However, for reasons outlined in Chapter 3, this was inappropriate and hence, an alternative was required. An approach to rescaling has been suggested recently in a condition-specific context (Ratcliffe, et al., 2009). However, this technique was reliant on attaching a value on the worst possible health state within the instrument, which was derived elsewhere. As described in Chapter 3, it was possible to rescale the mean score for each of the 18,000 health states described by the SF-6D. For all states, the utility weight that was attached to them was therefore

$$Utility_Weight = (\alpha + \beta X') / \alpha \quad \text{Equation 75}$$

As noted in Chapter 3, this is equivalent to the ratio of the marginal utility of extra time in the impaired health state, divided by the marginal utility of extra time in full health.

In situations in which the utility associated with time was assumed to be linear, this produces a QALY weight for each health state independent of the time variable *LIFE*, with full health anchored at 1. The explanation for this anchoring is that level one in each dimension of the SF-6D was omitted so the $\sum \beta X'$ term is equal to zero, meaning Equation (75) collapses to α / α . Note that Equation (75) allows the possibility of negative QALY weights if $\sum \beta X'$ exceeds the coefficient on life expectancy. For situations in which life expectancy was estimated as a non-linear function, QALY weights have to be estimated for a particular time point, and would be estimated as

$$\frac{\alpha + \beta X' + 2TIME(\rho + \tau X')}{\alpha + 2TIME(\rho + \tau X')} \quad \text{Equation 76}$$

In the analysis presented in this chapter, this was not done; however, the QALY weights using this non-linear utility function can be estimated using the results presented in Table 23.

To estimate confidence intervals around the QALY weights, the *wtp* command in STATA was employed with the default delta method (Hole, 2007b), involving a first-order Taylor expansion around the mean value of the variables, and then calculation of the variance of the resulting expression. The delta method has been shown to perform well, and to provide similar results to other competing approaches such as the Fieller or the Krinsky Robb methods (Hole, 2007b).

Additional sub-group analysis

Conventional valuation of generic health states for use in economic evaluation is focused on the mean respondent. This is appropriate within the convention that the value attributed to a health state is a societal one. However, it is useful to investigate

whether responses differ based on observable characteristics of the respondent. This is useful for two reasons. Firstly, it will identify the importance of using a balanced panel; if respondents do not differ in predictable ways, it is relatively less important that a balanced panel is used. The second reason is that it is intrinsically interesting to explore the degree to which people agree with the mean respondent.

The approach to doing this is to apply the RE probit (Model A) to sub-groups of the respondent space, making the assumption that the utility function is linear with respect to time. The reason I selected this random-effects probit rather than the more relaxed specifications is three-fold. Firstly, I want to investigate a variety of ways of dividing the respondents into groups and computational time becomes an issue if G-MNL or mixed logit models are employed. Secondly, and more importantly, the purpose of the sub-group analysis is not to investigate heterogeneity within a group of people with the same characteristics, but between groups of people with different characteristics. Finally, specifications beyond the RE probit estimate an increasingly large number of coefficients. Thus identifying patterns of statistically significant coefficients is difficult as chance would cause a number to be statistically significant (i.e. each unimportant coefficient has a probability equal to the chosen level of significance of being significant).

As stated previously, a range of demographic information was collected from respondents. Those that I selected for sub-group analyses were gender, age (i.e. above and below median age in the sample), education, studying status, gross household income, marital status, number of children, self-described health (within a 5-level Likert scale), whether the individual has a chronic condition, and whether they are currently employed. Where possible, the sample was split into approximately half, defined by each of these characteristics in turn. The utility function of alternative j for individual i with or without demographic characteristic c in scenario s is

$$U_{icsj} = X'_{isj}(\beta_i + \beta_c) + v_i + \varepsilon_{isj}, \quad \text{Equation 77}$$

where v_i is a person-specific error term ensuring that choices made by an individual are correlated, and β_c is a slope dummy testing for differences in attitudes towards health gain based on the characteristic c .

Each possible c was run separately. This was then compared with a pooled model, and the importance of the c terms were investigated using Information Criteria as in the main model evaluation section. Additionally, likelihood ratio tests were run. The results are presented graphically by rescaling the results from each sub-group such that the coefficient on duration is set to 1 (thus scale is corrected for). If, once the adjusted values are generated for two mutually exhaustive sub-groups, there is difference between the two for a particular level of a particular dimension, it means that the amount of life expectancy that an individual is willing to sacrifice to move to full health in that dimension differs.

Results

1,634 people entered the survey and were eligible to participate. Of these, 110 were removed because they exceeded the quota of respondents. Thus, they opened the link to the survey, but were immediately excluded. Of the remaining 1,524, 17 stated they were unwilling to participate, and 369 exited during the description of the task or before answering the first choice set. One hundred and twenty one respondents answered some of the choice tasks. Of the remaining 1,017 who responded to all choice sets, 13 failed to provide complete demographic information; this group was included in the analysis set. The 121 individuals completing some but not all of the choice sets were excluded from the analysis set. The reason for this decision was that the decision to drop out is likely to indicate the data from these individuals would show greater variability, and their responses were not required to ensure a large enough sample. The counter-argument is that the flexible approach to modelling scale used in this work would capture this, but at the time the trade-off was not considered valuable. The characteristics of the sample and its comparability to the general Australian population are outlined in Table 21.

Table 21: Representativeness of SF-6D DCE Sample

Characteristic	Value / Range	Sample	Population ²
Gender	Female	58.55%	56.09%
Age (years)	16-29	17.90%	21.33%
	30-44	19.00%	23.98%
	45-59	33.40%	22.40%
	60-74	28.20%	14.00%
	75+	1.50%	18.29%
Highest level of education	Primary	7.52%	40.51%
	Secondary	36.40%	20.00%
	Trade certificate	34.92%	22.24%
	Bachelor's degree or above	21.17%	17.26%
Gross household income ¹	<\$20,000	15.88%	15.77%
	\$20,000 - \$40,000	27.37%	23.02%
	\$40,001 - \$60,000	21.51%	17.64%
	\$60,001 - \$80,000	14.98%	13.87%
	\$80,001 - \$100,000	8.22%	11.03%
	\$100,001 +	12.05%	18.67%

¹ 15 individuals chose to not disclose income

² All data sourced from ABS (Australian Bureau of Statistics, 2006a; Australian Bureau of Statistics, 2002; Australian Bureau of Statistics, 2005; Australian Bureau of Statistics, 2007)

Thus, the respondents are over-representative of the middle age brackets, and are better educated than the general population. The importance of this will be investigated in the section concerned with sub-group analysis. The self-assessed health (as described by the SF-6D) of the individuals completing the survey is outlined in Table 22.

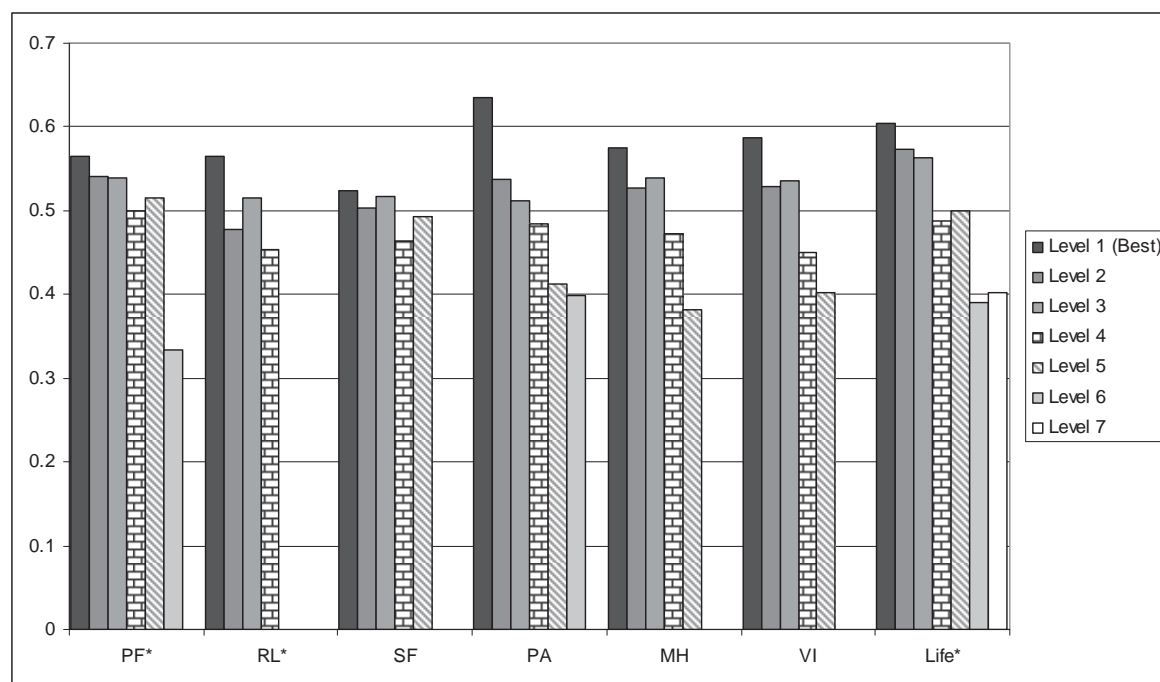
Table 22: Sample SF-6D Health (n=1,017)

Characteristic	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Physical Functioning	27.04%	37.07%	19.57%	13.57%	1.67%	1.08%
Role Limitation	54.38%	29.11%	7.57%	8.95%		
Social Functioning	50.74%	23.99%	17.11%	6.19%	1.97%	
Pain	29.20%	32.06%	18.09%	8.95%	7.77%	3.93%
Mental Health	21.14%	46.80%	22.81%	7.96%	1.28%	
Vitality	4.03%	37.76%	33.92%	20.55%	3.74%	

Marginal frequencies

The marginal frequencies for each level are shown in Figure 16.

Figure 16: SF-6D Dimension / Level Marginal Frequencies



† The levels in this figure refer to the levels of the SF-6D, with the exception of Life (Expectancy), for which the levels in the figure are 20 years, 16 years, 12 years, 8 years, 4 years, 2 years and 1 year, labelled as Level 1-7 respectively.

* The starred levels are those that are not strictly monotonic (and hence not have a monotonic relationship displayed in the marginal frequencies)

Most dimensions show the expected monotonic relationship as the dimension moves to increasingly poorer health. The slight breakdown of this pattern in some dimensions (Physical Functioning, Role Limitation and Life Expectancy) can potentially be explained by arguing that these dimensions are not strictly monotonic in

their construction. This issue was discussed in Chapter 2. The dimensions in which the gradient is steepest (Pain, Mental Health, Vitality) are those that most impacted on the decision to prefer either option in the DCE; this pattern is therefore expected to be reflected in the econometric models.

Across all choice sets, the death option was selected as the best of the three in 2.7% of the choice sets, and as the worst in 57.4%. As noted previously in this chapter, the preference for health states relative to immediate death is not used in estimating models; nevertheless, it is noteworthy that in 42.6% of choice sets, the respondent was willing to state that death was preferable to at least one of the choice sets, a result which suggests the floor effect seen in the UK algorithm of Brazier is not reflective of population attitudes to the poorest levels of health in the SF-6D.

The results from Models A-D are presented in Table 23.

Table 23: Results From Models A-D

Dimension	Level	Model A	Model B	Model C	Model D
Duration	Linear	0.1908 (0.0054)	0.1915 (0.0053)	0.1936 (0.0055)	0.4674 (0.0189)
	2	-0.0090 (0.0028)	-0.0091 (0.0028)	-0.0090 (0.0028)	-0.0250 (0.0114)
	3	-0.0147 (0.0027)	-0.0150 (0.0027)	-0.0144 (0.0027)	-0.0214 (0.0101)
	4	-0.0264 (0.0027)	-0.0266 (0.0027)	-0.0269 (0.0027)	-0.0784 (0.0112)
	5	-0.0268 (0.0029)	-0.0274 (0.0028)	-0.0266 (0.0029)	-0.0634 (0.0115)
	6	-0.0563 (0.0029)	-0.0567 (0.0028)	-0.0537 (0.0030)	-0.1332 (0.0121)
Role Limitation (RL) x Duration	2	-0.0181 (0.0025)	-0.0172 (0.0024)	-0.0164 (0.0026)	-0.0412 (0.0102)
	3	-0.0132 (0.0023)	-0.0126 (0.0022)	-0.0124 (0.0023)	-0.0102 (0.0093)
	4	-0.0226 (0.0024)	-0.0218 (0.0023)	-0.0196 (0.0027)	-0.0250 (0.0095)
	5	-0.0048 (0.0028)	-0.0052 (0.0028)	-0.0040 (0.0028)	-0.0684 (0.0116)
Social Functioning (SF) x Duration	3	-0.0059 (0.0025)	-0.0065 (0.0024)	-0.0064 (0.0025)	-0.0350 (0.0106)
	4	-0.0215 (0.0024)	-0.0217 (0.0024)	-0.0211 (0.0024)	-0.0925 (0.0108)
	5	-0.0239 (0.0027)	-0.0246 (0.0026)	-0.0215 (0.0028)	-0.0849 (0.0106)
	2	-0.0151 (0.0029)	-0.0154 (0.0029)	-0.0157 (0.0029)	-0.0747 (0.0121)
	3	-0.0329 (0.0027)	-0.0333 (0.0026)	-0.0334 (0.0027)	-0.1023 (0.0107)
Pain (PA) x Duration	4	-0.0390 (0.0029)	-0.0390 (0.0029)	-0.0386 (0.0029)	-0.0686 (0.0116)
	5	-0.0553 (0.0027)	-0.0539 (0.0024)	-0.0551 (0.0027)	-0.1349 (0.0109)
	6	-0.0520 (0.0030)		-0.0497 (0.0031)	-0.1920 (0.0132)
	2	-0.0115 (0.0024)	-0.0117 (0.0024)	-0.0110 (0.0024)	-0.0338 (0.0101)
	3	-0.0145 (0.0024)	-0.0149 (0.0024)	-0.0137 (0.0024)	-0.0237 (0.0103)
	4	-0.0356 (0.0026)	-0.0360 (0.0025)	-0.0351 (0.0026)	-0.1071 (0.0113)
Mental Health (MH) x Duration	5	-0.0521 (0.0025)	-0.0523 (0.0025)	-0.0494 (0.0026)	-0.1412 (0.0099)
	2	-0.0020 (0.0024)	-0.0021 (0.0024)	-0.0023 (0.0024)	-0.0501 (0.0120)
	3	-0.0089 (0.0027)	-0.0091 (0.0027)	-0.0107 (0.0028)	-0.0741 (0.0104)
	4	-0.0406 (0.0026)	-0.0406 (0.0026)	-0.0422 (0.0027)	-0.0987 (0.0108)
	5	-0.0480 (0.0028)	-0.0483 (0.0028)	-0.0460 (0.0029)	-0.1659 (0.0106)
Vitality (VI) x Duration	Most			-0.0072 (0.0029)	

Duration ²	Quadratic				
Duration ² x PF	2				-0.0197 (0.0012)
	3				0.0021 (0.0007)
	4				0.0008 (0.0006)
	5				0.0031 (0.0007)
	6				0.003 (0.0008)
					0.0061 (0.0008)
Duration ² x RL	2				0.0020 (0.0007)
	3				-0.0003 (0.0006)
	4				0.0010 (0.0006)
Duration ² x SF	2				0.0043 (0.0007)
	3				0.0024 (0.0007)
	4				0.0047 (0.0007)
	5				0.0042 (0.0007)
					0.0042 (0.0008)
Duration ² x PA	2				0.0052 (0.0007)
	3				0.0025 (0.0008)
	4				0.0058 (0.0007)
	5				0.0094 (0.0008)
	6				0.0016 (0.0006)
					0.0010 (0.0006)
Duration ² x MH	2				0.0049 (0.0007)
	3				0.0063 (0.0006)
	4				0.0031 (0.0008)
	5				0.0042 (0.0007)
					0.0042 (0.0007)
Duration ² x VI	2				0.0084 (0.0007)
	3				
	4				
	5				

	constant	0.0128 (0.0111)	0.0128 (0.0111)	-0.0075 (0.0109)	-0.0057 (0.0110)
	ρ	0.0041 (0.0056)	0.0040 (0.0056)	0.0000 (0.0000)	0.0000 (0.0000)
	Log likelihood	-8972	-8973	-8970	-8667
	AIC	18000	18000	17998	17385
	BIC	18214	18206	18219	17789

* Coefficients in **bold** are statistically significant at the 1% level

Before looking at the patterns in these results, I want to confirm that the choice of a probit specification over a logit does not impact on inferences from these results. The logit results for Models A and D are presented in Appendix 5, and summarised graphically in Figure 17 and Figure 18.

Figure 17: RE probit and logit coefficients (Model A)

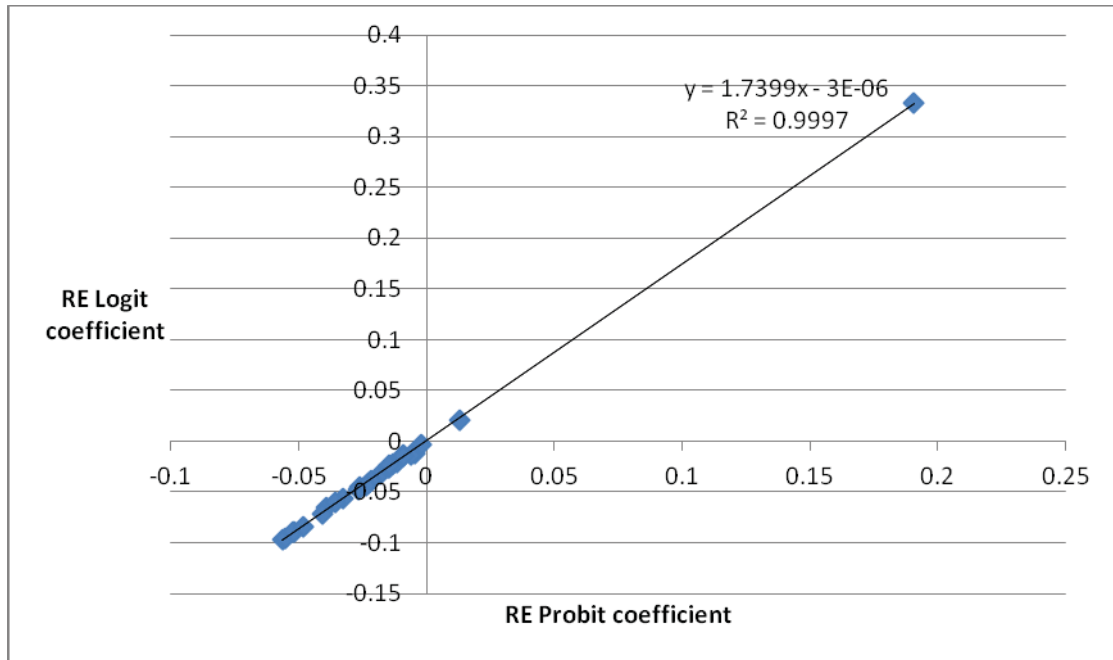
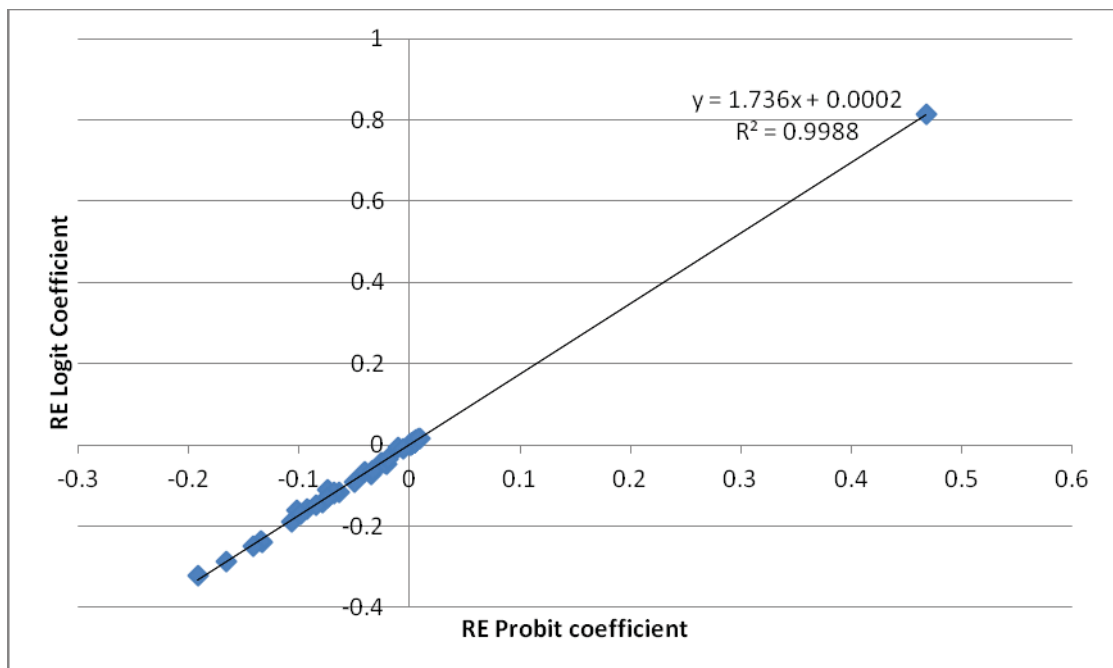


Figure 18: RE probit and logit coefficients (Model D)



Under either the linear or non-linear utility function, the logit coefficients are all absolutely larger than those in the probit. Due to the highly correlated coefficients under the two models, the inferences using either a logit or a probit are likely to be very similar. Therefore, for the base case results, I consider only the RE probit results.

Model B combines levels 5 and 6 for Pain. This combination does not cause any new non-monotonic orderings, and because the unadjusted coefficients were close, has very little impact on log likelihood (it does improve BIC as fewer coefficients are estimated; however, it would not be appropriate to combine other pairs of consecutive levels to further improve BIC, as there is no good a priori reason for the ordering reflected in the unadjusted results to be over-ruled). It is these adjusted regression results (with levels 5 and 6 of pain combined but no others) that are used in the estimation of QALY weights for the SF-6D.

Introducing the *MOST* term in Model C has a small impact on log-likelihood and Information Criteria. The coefficient is negative suggesting that the first dimension to move to the worst level has an extra disutility than subsequent dimensions. This result is reflected in other algorithms (Dolan, 1997; Viney, et al., 2011b).

Model D leads to a significant improvement in model fit. The quadratic term on duration is highly statistically significant. In addition, the coefficients on the other quadratic terms (i.e. those interacted with levels of SF-6D dimensions) are frequently statistically significant and mostly positive. The decision regarding which model is preferred is difficult. The models which impose a linear utility function with respect to time produce more straightforward QALY weights, in that they are independent of time. However, the models which relax this assumption perform better under the Information Criteria, and have 11 statistically significant coefficients.

Base case utility weights for the SF-6D

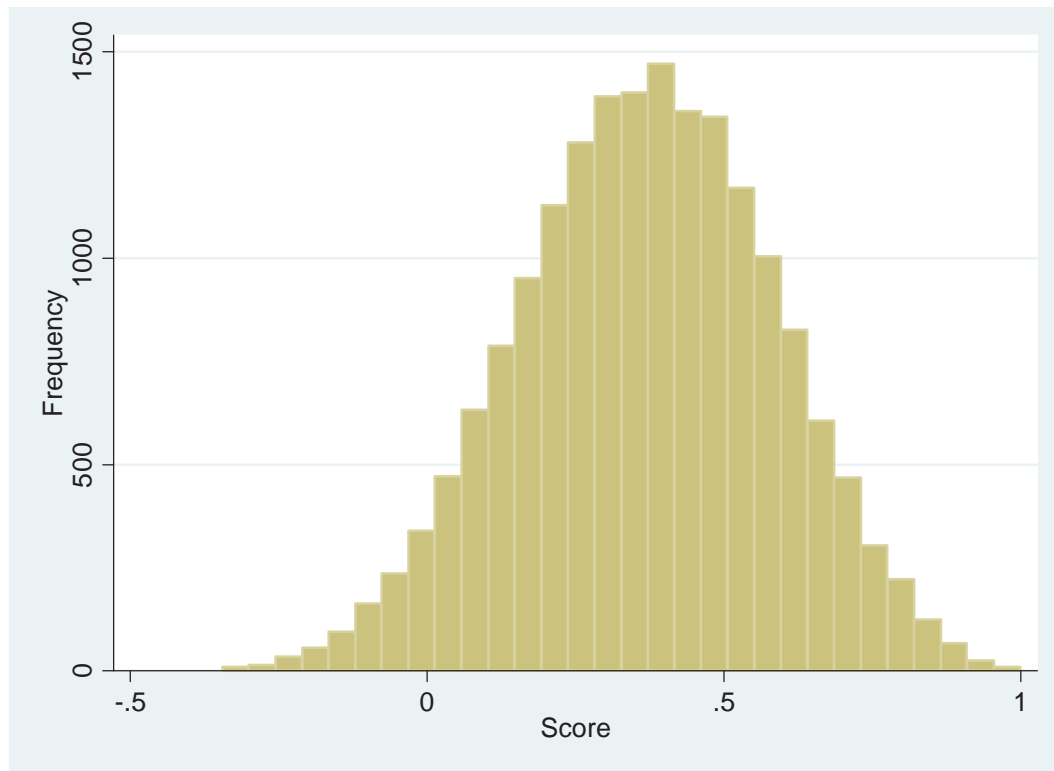
In Chapter 3, the methods for converting regression results into utility weights were discussed. If the linear utility function is assumed, and the RE-probit results are utilised, the weights (and 95% confidence intervals) to generate the scores for each of the SF-6D health states are presented in Table 24.

Table 24: Base case QALY algorithm

	Model B	Model C
Level	QALY decrement (95% CI)	QALY decrement (95% CI)
PF2	0.0474 (0.0194-0.0754)	0.0464 (0.0186-0.0741)
PF3	0.0776 (0.0509-0.1043)	0.0744 (0.0479-0.1009)
PF4	0.1390 (0.1139-0.1642)	0.1389 (0.1141-0.1638)
PF5	0.1428 (0.1166-0.1690)	0.1374 (0.1111-0.1638)
PF6	0.2961 (0.2686-0.3236)	0.2774 (0.2470-0.3077)
RL2	0.0896 (0.0657-0.1136)	0.0847 (0.0581-0.1114)
RL3	0.0658 (0.0430-0.0886)	0.0639 (0.0403-0.0874)
RL4	0.1142 (0.0905-0.1379)	0.1011 (0.0731-0.1291)
SF2	0.0271 (-0.0007-0.0550)	0.0204 (-0.0077-0.0486)
SF3	0.0337 (0.0094-0.0580)	0.0331 (0.0084-0.0577)
SF4	0.1131 (0.0906-0.1356)	0.1089 (0.0864-0.1313)
SF5	0.1280 (0.1040-0.1521)	0.1112 (0.0844-0.1380)
PA2	0.0807 (0.0525-0.1089)	0.0811 (0.0531-0.1091)
PA3	0.1742 (0.1482-0.2001)	0.1727 (0.1470-0.1984)
PA4	0.2038 (0.1757-0.2318)	0.1992 (0.1712-0.2272)
PA5	0.2814 (0.2580-0.3049)	0.2847 (0.2569-0.3126)
PA6	0.2814 (0.2580-0.3049)	0.2568 (0.2262-0.2874)
MH2	0.0616 (0.0375-0.0857)	0.0566 (0.0325-0.0806)
MH3	0.0780 (0.0537-0.1022)	0.0708 (0.0463-0.0953)
MH4	0.1881 (0.1634-0.2129)	0.1812 (0.1564-0.2061)
MH5	0.2731 (0.2489-0.2972)	0.2552 (0.2279-0.2825)
VI2	0.0110 (-0.0133-0.0352)	0.0117 (-0.0122-0.0357)
VI3	0.0476 (0.0206-0.0746)	0.0553 (0.0279-0.0827)
VI4	0.2117 (0.1887-0.2347)	0.2181 (0.1950-0.2412)
VI5	0.2521 (0.2284-0.2759)	0.2377 (0.2118-0.2636)
MOST		0.0373 (0.0083-0.0663)

To value a health state, the relevant weights are subtracted from 1. As an example, to value health state 321234 under Model B, the value is $1 - (0.0776 + 0.0896 + 0.0807 + 0.0780 + 0.2117) = 0.4624$. The scores for the 18,000 health states within the SF-6D instrument under the corrected RE probit result are presented in Figure 19.

Figure 19: Distribution of SF-6D health states (corrected utility function 1, random-effects probit)



It is clear from this histogram that the distribution of weights under the Australian DCE algorithm is very different to that in the original UK weights. Most notably, the Australian algorithm allows weights below zero, while the UK algorithm has a floor effect at 0.3. A more thorough comparison of utility weights under the two algorithms is provided in the discussion section of this chapter.

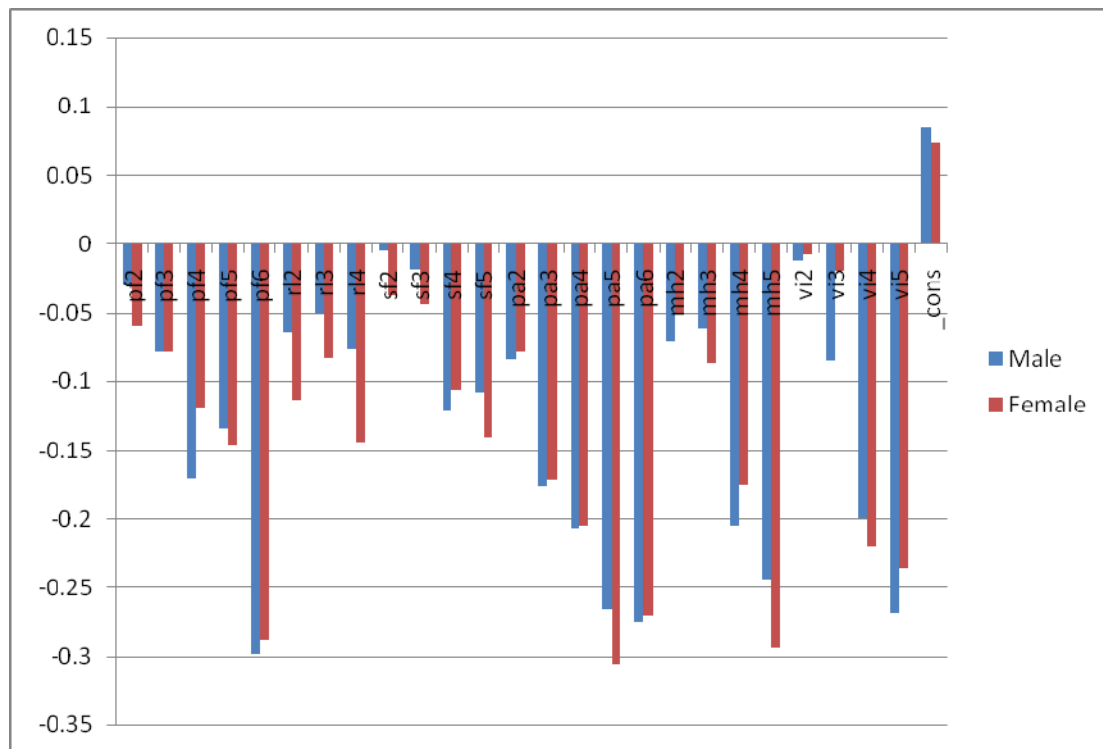
The next step in the chapter is to consider whether these results differ predictably based on observed respondent characteristics.

Sub-group analysis

Gender

The results dividing the sample by gender suggest there is little difference in responses. Note that the coefficients are normalised such that the coefficient for duration is set to be 1 (but not shown in the figure), thus coefficients are comparable across samples. If these adjusted coefficients are generally higher for a sub-group, this suggests that group places less emphasis on the dimension which has been normalised. The full tabulated results are provided in Appendix 6.

Figure 20: Sub-Group Analysis Results (Gender of respondent)



The model comparison information for the gender sub-group analysis is presented in Table 25. Note that the constrained specification is that which pools male and female responses, while the unconstrained specification allows them to differ.

Table 25: Information Criteria (Gender Sub-Group Analysis)

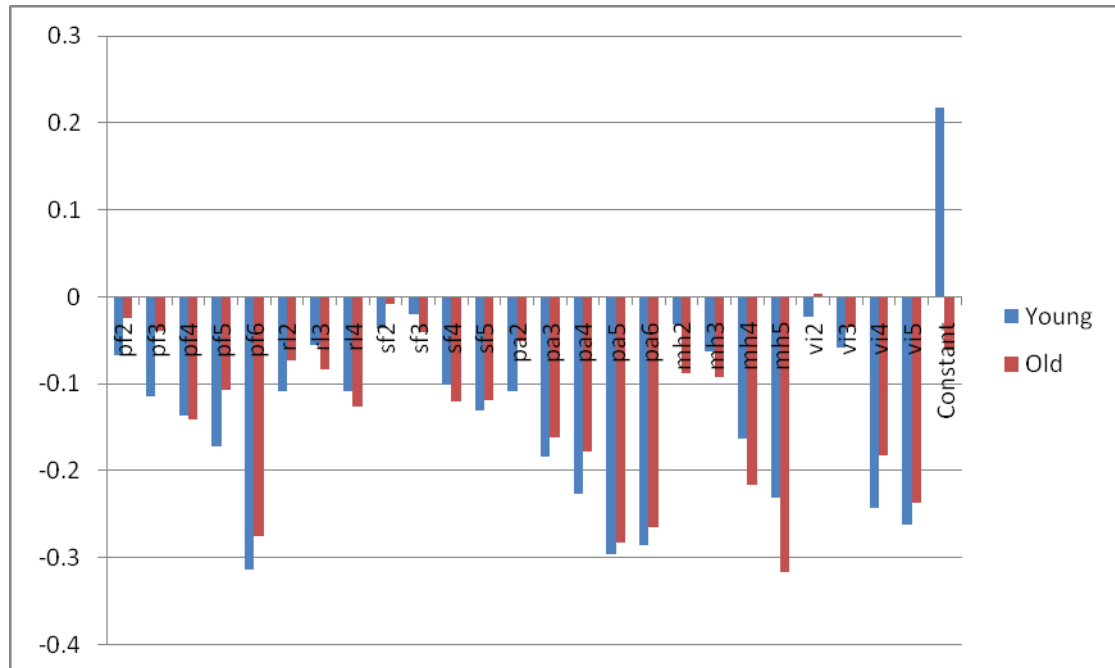
Model	Observations	Log likelihood	Degrees of freedom	AIC	BIC
Male and female pooled	15090	-8879	28	17814	18027
Unconstrained	15090	-8859	54	17827	18238

Both the AIC and BIC suggest that allowing coefficients to differ based on the gender of the respondent is not necessary. However, the likelihood ratio test, which tests whether it is appropriate to nest the restricted model (i.e. in which no allowance is made for the gender of the respondent) within the more unrestricted model suggests it may not be appropriate to pool the results between the genders (lr $\chi^2(26)=39.5$; $p=0.0437$). One reason for accepting poolability between genders is that it is difficult to determine a clear pattern of differences between male and female respondents. Of the 6 dimensions, only Role Limitation has one of the genders (female) considering each level more serious than males do.

Age

The median age of respondents was 51.5 years. Replicating the sub-group analysis allowing coefficients to differ for those older and younger than this median are presented in Appendix 7.

Figure 21: Sub-Group Analysis Results (Age of Respondent)



As described previously, these samples can be directly compared because the coefficient on duration has been normalised to be one (analogous to a willingness to pay calculation). In this instance, there is a possible pattern in that the respondents below median age place relatively greater emphasis on physical functioning, pain and vitality, while those above median age place relatively greater emphasis on mental health and perhaps also role limitation.

From a survey design perspective, it is interesting to note that the constant is notably more positive in the younger cohort. Thus, it is this group which is systematically more likely to select choice set A. As I randomised which option was in position 1 and position 2, this will not systematically bias the regression results; however, it remains unexplained, The information criteria for the age sub-group analysis are presented in Table 26.

Table 26: Information Criteria (Age Sub-Group Analysis)

Model	Observations	Log	Degrees	AIC	BIC
-------	--------------	-----	---------	-----	-----

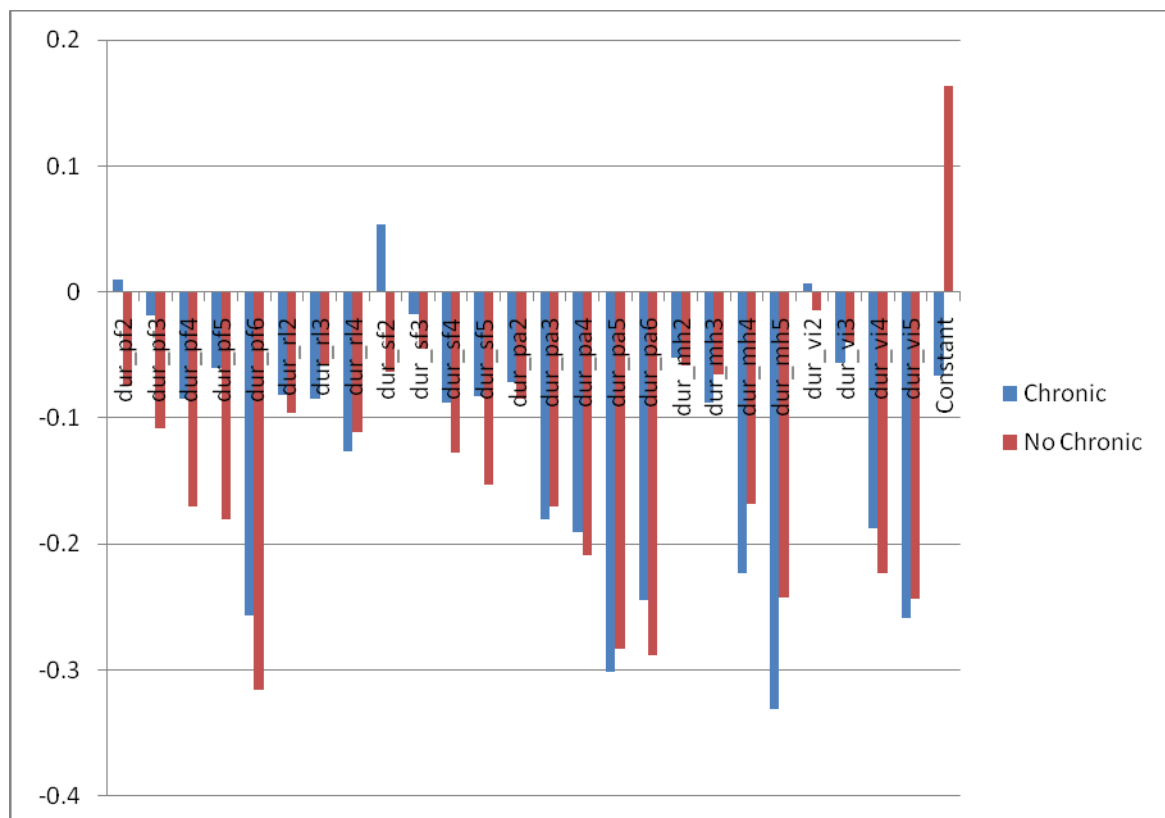
		likelihood	of freedom		
Constrained	15150	-8913	28	17882	18096
Unconstrained	15150	-8894	54	17896	18308

As with the gender division, the AIC suggests the consideration of age of respondent in the analysis does little regarding model fit. Equally, the BIC suggests the constrained model is the more appropriate. As with the case of gender, the likelihood ratio test suggests that it is questionable whether the groups can be pooled ($\text{lr } \chi^2(26)=38.51; p=0.0542$). Relative to the gender case, it is easier to discern patterns within dimensions when the sample is sub-divided by age. In three of the six dimensions, older people value all levels either better than younger people do (Vitality, Pain), or worse than younger people do (Mental Health).

Chronic Conditions

Of those who provided information regarding chronic conditions (as the respondent was allowed to refuse to answer) 39.3% of the sample defined themselves as having a chronic condition. Thus, the RE probit was re-run allowing coefficients to differ based on this variable. The results were again normalised and graphed in Figure 22.

Figure 22: Sub-Group Analysis Results (Chronic Conditions)



Respondents with chronic conditions appear to place relatively greater emphasis on mental health, and relatively less emphasis on physical functioning and social functioning. Whether an individual has a chronic condition is likely to be correlated with their age. While possible to run sub-group analysis accounting for both issues, the increasingly small populations in each sub-group and the proliferation of coefficients make this analysis unhelpful. For example, running these together would require four coefficients to be estimated for level of each dimension (e.g. for Physical Functioning 2, the analysis would require the following terms to be estimated: 1) Duration x PF2; 2) Duration x PF2 x Chronic Only; 3) Duration x PF2 x Old Only; and 4) Duration x PF2 x (Chronic and Old).). As respondent characteristics are not independent, the number of observations driving each of these coefficients would differ, and be small for some).

The Information Criteria for these results are presented in Table 27.

Table 27: Information Criteria (Chronic Conditions Sub-Group Analysis)

Model	Observations	Log likelihood	Degrees of freedom	AIC	BIC
Constrained	15060	-8864	28	17783	17997
Unconstrained	15060	-8841	54	17791	18202

As before, both Information Criteria suggest adopting the constrained model (and hence accepting the pooling of data from those with and without a chronic condition. Unlike the previous two analyses, the LR test unequivocally rejects poolability between people with and without a chronic condition ($\ln \chi^2(26)=46.44$; $p=0.0081$). In this case, people with a chronic condition value all levels better than people without a chronic condition do (Physical and Social Functioning).

Heterogeneity modelling

Utility Function A

The results for the various approaches to modelling heterogeneity are now presented. The results for Utility Function A under each of the approaches are presented in Table 28.

Table 28: Heterogeneity Modelling Specification Results (Utility Model A)

Mean Model	Clogit		Scale MNL		Uncorrelated coefficients		Correlated coefficients‡	
	AI	A1	A2	A2	Mixed logit	G-MNL	Mixed logit	G-MNL
Duration	0.333 (0.010)***	0.528 (0.029)***	0.006 (0.008)	0.453 (0.014)***	2.014 (0.244)***	0.501 (0.017)***	1.903 (0.239)***	
PF2	-0.014 (0.005)***	0.006 (0.008)	0.006 (0.008)	-0.016 (0.006)***	-0.039 (0.032)	-0.017 (0.007)**	-0.028 (0.016)*	
PF3	-0.026 (0.005)***	-0.036 (0.008)***	-0.036 (0.008)***	-0.032 (0.006)***	-0.117 (0.027)***	-0.029 (0.006)***	-0.106 (0.021)***	
PF4	-0.045 (0.005)***	-0.062 (0.008)***	-0.062 (0.008)***	-0.058 (0.006)***	-0.233 (0.039)***	-0.063 (0.006)***	-0.224 (0.034)***	
PF5	-0.047 (0.005)***	-0.061 (0.008)***	-0.061 (0.008)***	-0.062 (0.006)***	-0.242 (0.038)***	-0.067 (0.007)***	-0.227 (0.031)***	
PF6	-0.096 (0.005)***	-0.127 (0.009)***	-0.127 (0.009)***	-0.125 (0.006)***	-0.490 (0.061)***	-0.143 (0.008)***	-0.476 (0.058)***	
RL2	-0.032 (0.004)***	-0.047 (0.007)***	-0.047 (0.007)***	-0.039 (0.006)***	-0.131 (0.021)***	-0.038 (0.006)***	-0.155 (0.024)***	
RL3	-0.023 (0.004)***	-0.031 (0.006)***	-0.031 (0.006)***	-0.025 (0.005)***	-0.071 (0.018)***	-0.022 (0.006)***	-0.071 (0.020)***	
RL4	-0.042 (0.004)***	-0.070 (0.008)***	-0.070 (0.008)***	-0.051 (0.005)***	-0.204 (0.028)***	-0.057 (0.006)***	-0.210 (0.030)***	
SF2	-0.012 (0.005)**	-0.040 (0.008)***	-0.040 (0.008)***	-0.013 (0.006)**	-0.076 (0.026)***	-0.012 (0.006)*	-0.069 (0.019)***	
SF3	-0.013 (0.004)***	-0.033 (0.007)***	-0.033 (0.007)***	-0.019 (0.005)***	-0.107 (0.021)***	-0.021 (0.006)***	-0.092 (0.018)***	
SF4	-0.039 (0.004)***	-0.065 (0.007)***	-0.065 (0.007)***	-0.050 (0.005)***	-0.204 (0.028)***	-0.052 (0.005)***	-0.185 (0.027)***	
SF5	-0.043 (0.005)***	-0.078 (0.009)***	-0.078 (0.009)***	-0.061 (0.006)***	-0.251 (0.034)***	-0.063 (0.007)***	-0.263 (0.040)***	
PA2	-0.026 (0.005)***	-0.034 (0.007)***	-0.034 (0.007)***	-0.041 (0.006)***	-0.119 (0.024)***	-0.046 (0.007)***	-0.159 (0.025)***	
PA3	-0.056 (0.005)***	-0.086 (0.007)***	-0.086 (0.007)***	-0.071 (0.006)***	-0.252 (0.034)***	-0.076 (0.006)***	-0.297 (0.040)***	
PA4	-0.066 (0.005)***	-0.100 (0.008)***	-0.100 (0.008)***	-0.084 (0.006)***	-0.277 (0.036)***	-0.091 (0.007)***	-0.327 (0.042)***	
PA5	-0.096 (0.005)***	-0.159 (0.011)***	-0.159 (0.011)***	-0.125 (0.006)***	-0.473 (0.056)***	-0.138 (0.007)***	-0.522 (0.063)***	
PA6	-0.090 (0.005)***	-0.140 (0.009)***	-0.140 (0.009)***	-0.123 (0.007)***	-0.508 (0.067)***	-0.144 (0.009)***	-0.549 (0.072)***	
MH2	-0.021 (0.004)***	-0.027 (0.007)***	-0.027 (0.007)***	-0.022 (0.005)***	-0.125 (0.021)***	-0.022 (0.005)***	-0.126 (0.022)***	
MH3	-0.023 (0.004)***	-0.016 (0.006)**	-0.016 (0.006)**	-0.017 (0.005)***	-0.071 (0.023)***	-0.018 (0.006)***	-0.041 (0.018)**	
MH4	-0.060 (0.004)***	-0.087 (0.008)***	-0.087 (0.008)***	-0.080 (0.006)***	-0.349 (0.046)***	-0.091 (0.006)***	-0.367 (0.049)***	
MH5	-0.089 (0.004)***	-0.128 (0.008)***	-0.128 (0.008)***	-0.113 (0.006)***	-0.461 (0.060)***	-0.131 (0.007)***	-0.481 (0.065)***	
VI2	-0.003 (0.004)	-0.002 (0.006)	-0.002 (0.006)	-0.004 (0.005)	-0.021 (0.020)	-0.007 (0.005)	-0.015 (0.014)	
VI3	-0.015 (0.005)***	-0.021 (0.007)***	-0.021 (0.007)***	-0.020 (0.005)***	-0.085 (0.019)***	-0.025 (0.006)***	-0.093 (0.020)***	
VI4	-0.071 (0.005)***	-0.106 (0.008)***	-0.106 (0.008)***	-0.081 (0.005)***	-0.307 (0.038)***	-0.091 (0.006)***	-0.328 (0.044)***	

VI5	-0.084 (0.005)***	-0.121 (0.009)***	-0.104 (0.006)***	-0.397 (0.051)***	-0.116 (0.007)***	-0.437 (0.057)***
τ		1.054 (0.054)		1.627 (0.087)		1.702 (0.089)
Standard Deviation [†]						
Duration			0.166 (0.009)***	0.705 (0.099)***	0.191 (0.011)***	0.606 (0.078)***
PF6			0.040 (0.011)***	0.269 (0.036)***	0.080 (0.009)***	0.268 (0.040)***
RL4			0.035 (0.008)***	0.188 (0.030)***	0.064 (0.007)***	0.175 (0.023)***
SF5			0.004 (0.024)	0.173 (0.030)***	0.038 (0.009)***	0.120 (0.020)***
PA6			0.064 (0.010)***	0.262 (0.040)***	0.085 (0.010)***	0.239 (0.035)***
MH5			0.075 (0.007)***	0.244 (0.037)***	0.090 (0.007)***	0.259 (0.036)***
VI5			0.060 (0.008)***	0.242 (0.035)***	0.076 (0.008)***	0.299 (0.040)***
γ				-0.104 (0.026)***		-0.014 (0.015)
Log likelihood	-8943	-8738	-8674	-8411	-8614	-8368
Degrees of freedom	26	27	33	36	54	56
AIC	17938	17530	17415	16984	17337	16849
BIC (n= individuals)	18066	17530	17574	17069	17598	17120
BIC (n = observations)	18136	17755	17690	17193	17787	17315

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

[†] STATA reports standard deviations as both positive and negative, but notes that the sign is irrelevant. Therefore, the absolute values are presented here

[‡] Here, the standard deviations for each of the random variables were generated to allow some comparability with other models generating standard deviation figures. These were generated using either the STATA *mixlcv, sd* or the *gmmicov, sd* command. The variance-covariance matrices for Models A5 and A6 are provided in Table 29 and Table 30 respectively.

The conditional logit results (Model A1) suggest that utility increases in life expectancy, and reduces as the health state moves away from full health. The monotonicity implicit in some of the SF-6D dimensions is violated in two instances: Social Functioning Level 3 is valued as being a smaller decrement to utility than Social Functioning Level 2; and similarly Pain Level 6 is considered less of a decrement than Level 5. While the absolute size of the coefficient on Role Limitation Level 3 is smaller than that on Level 2, I have previously argued that the values over these levels need not be monotonic, and therefore this result is not a violation of expected orderings.

Physical Functioning level 2 loses statistical significance under the S-MNL, and Social Functioning level 2 gains significance at the 1% level. All other conclusions regarding statistical significance remain the same as those presented in Model A1. However, the impact of allowing for scale on model fit is noteworthy. The log-likelihood improves by 205 points. As discussed in Chapter 3, the inclusion of a scale parameter alone is a highly parsimonious addition to the simple conditional logit model. This is reflected in the improved Information Criteria figures under Model A2 than under A1.

The mixed logit results (A3) suggest there is considerable heterogeneity between respondents. Six of the seven random coefficients are statistically significant at the 1% level. The mean responses show a similar monotonic pattern to that identified in the conditional logit and S-MNL results. While there are six additional parameters relative to the S-MNL (and seven relative to the conditional logit), the AIC and BIC suggest that this extra estimation improves model fit. As previously argued, the value each respondent places on particular coefficients is likely to be highly correlated with values placed on other coefficients. Model A5 (which replicates Model A3 but allows correlation) suggests this might be the case: the AIC in Model A5 is superior, but the BIC (using either interpretation of the total n) shows the opposite pattern. It is likely therefore, that if the poorer levels of health are those most likely to demonstrate correlation, that allowing additional correlations is not justified given the equivocal findings for allowing correlations in this limited set.

The uncorrelated G-MNL model (A4) results suggest that considering both scale and preference heterogeneity in one model leads to an improvement in model fit relative

to considering only one, or neither. In both the correlated and independent coefficient cases, the signs of the coefficients are generally as expected, and the monotonic nature of the SF-6D is broadly reflected. The AIC and BIC are favourable, reflecting both the good model fit and the relative parsimony of the approach (in that, relative to the mixed logit presented as Model A3, there are only two additional parameters). Relative to Model A5, Model A4 performs very well, both with a better log-likelihood and fewer estimated coefficients. The γ term is very close to zero, suggesting that the scale term applies almost equally to the parameter coefficient β and the variance term η_i (this is termed by Fiebig (2010) as G-MNL-II). Indeed, in Model A4, the γ term is negative, suggesting that the scale term applies more to the variance terms than to the parameter coefficients.

Model A6 performs well relative to A5, suggesting that the consideration of scale leads to a considerable improvement in model fit. However, relative to Model A4, the Bayesian information criteria suggests that the additional model fit allowed by introducing correlated coefficients, does not justify the additional 20 degrees of freedom required (a pattern contradicted by the AIC). This equivocal pattern in terms of Information Criteria occurs despite there being a number of statistically significant terms in the correlation matrix for Model A6 presented in Table 30.

The variance-covariance matrices for Models A5 and A6 are presented in Table 29 and Table 30.

Table 29: Variance-Covariance Matrices for Model A5

	Duration	PF6	RL4	SF5	PA6	MH5	VI5
Duration	0.191 (0.011) ***						
PF6	-0.032 (0.008) ***	0.074 (0.009) ***					
RL4	-0.011 (0.006) *	0.039 (0.008) ***	0.049 (0.008) ***				
SF5	0.003 (0.007)	-0.002 (0.01)	-0.016 (0.009) *	-0.035 (0.009) ***			
PA6	-0.032 (0.008) ***	0.01 (0.011)	-0.002 (0.014)	-0.041 (0.015) ***	0.066 (0.012) ***		
MH5	-0.027 (0.007) ***	0.014 (0.011)	0.045 (0.011) ***	-0.045 (0.014) ***	-0.043 (0.015) ***	0.037 (0.016) **	
VI5	-0.011 (0.007)	0.013 (0.01)	-0.023 (0.01) **	0.029 (0.014) **	0(0.012)	0.062 (0.01) ***	0.016 (0.022)

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)
Standard errors noted in parentheses

Table 30: Variance-Covariance Matrices for Model A6

	Duration	PF6	RL4	SF5	PA6	MH5	VI5
Duration	0.606 (0.078) ***						
PF6	-0.057 (0.017) ***	0.261 (0.038) ***					
RL4	0.012 (0.014)	0.108 (0.017) ***	0.137 (0.019) ***				
SF5	0.000 (0.015)	-0.099 (0.017) ***	0.000 (0.013)	0.068 (0.017) ***			
PA6	-0.021 (0.014)	0.006 (0.016)	0.011 (0.015)	0.132 (0.026) ***	0.198 (0.030) ***		
MH5	0.002 (0.012)	0.055 (0.012) ***	0.135 (0.019) ***	0.152 (0.023) ***	-0.151 (0.025) ***	-0.003 (0.011)	
VI5	0.006 (0.009)	0.011 (0.010)	-0.113 (0.020) ***	-0.111 (0.018) ***	-0.098 (0.021) ***	-0.118 (0.018) ***	0.201 (0.027) ***

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)
Standard errors noted in parentheses

Utility function D

The corresponding results for Models D1-D6 (which differ from A1-A6 in that they allow for non-linearity of the utility function over time) are presented in Table 31.

Table 31: Heterogeneity Modelling Specification Results (Utility Model D)

	Clogit		Scale MNL		Uncorrelated coefficients		Correlated coefficients†	
	D1	D2	Mixed logit	D3	D4	D5	D6	
Mean								
Model								
Duration	0.815 (0.033)***	1.156 (0.060)***	0.943 (0.04)***	0.943 (0.04)***	2.218 (0.247)***	0.953 (0.041)***	2.798 (0.321)***	
Duration ²	-0.035 (0.002)***	-0.050 (0.003)***	-0.036 (0.003)***	-0.036 (0.003)***	-0.084 (0.010)***	-0.036 (0.003)***	-0.106 (0.013)***	
PF2	-0.045 (0.019)**	-0.079 (0.026)***	-0.034 (0.022)	-0.034 (0.022)	-0.168 (0.058)***	-0.032 (0.022)	-0.221 (0.065)***	
PF3	-0.046 (0.017)***	-0.102 (0.025)***	-0.032 (0.020)	-0.032 (0.020)	-0.164 (0.055)***	-0.029 (0.020)	-0.154 (0.056)***	
PF4	-0.141 (0.019)***	-0.223 (0.028)***	-0.137 (0.023)***	-0.137 (0.023)***	-0.427 (0.077)***	-0.137 (0.023)***	-0.444 (0.076)***	
PF5	-0.117 (0.020)***	-0.186 (0.029)***	-0.118 (0.023)***	-0.118 (0.023)***	-0.372 (0.076)***	-0.115 (0.024)***	-0.443 (0.081)***	
PF6	-0.240 (0.021)***	-0.341 (0.031)***	-0.252 (0.025)***	-0.252 (0.025)***	-0.626 (0.092)***	-0.263 (0.025)***	-0.714 (0.096)***	
RL2	-0.067 (0.017)***	-0.050 (0.023)**	-0.078 (0.020)***	-0.078 (0.020)***	-0.132 (0.042)***	-0.075 (0.021)***	-0.130 (0.044)***	
RL3	-0.007 (0.016)	0.024 (0.021)	-0.039 (0.018)**	-0.039 (0.018)**	-0.094 (0.042)**	-0.037 (0.018)**	-0.098 (0.047)**	
RL4	-0.045 (0.016)***	-0.049 (0.021)**	-0.064 (0.019)***	-0.064 (0.019)***	-0.148 (0.042)***	-0.069 (0.019)***	-0.198 (0.047)***	
SF2	-0.121 (0.020)***	-0.161 (0.027)***	-0.153 (0.023)***	-0.153 (0.023)***	-0.254 (0.050)***	-0.161 (0.024)***	-0.315 (0.059)***	
SF3	-0.071 (0.018)***	-0.127 (0.024)***	-0.078 (0.020)***	-0.078 (0.020)***	-0.201 (0.041)***	-0.078 (0.021)***	-0.236 (0.048)***	
SF4	-0.158 (0.018)***	-0.225 (0.026)***	-0.185 (0.021)***	-0.185 (0.021)***	-0.414 (0.060)***	-0.192 (0.022)***	-0.503 (0.073)***	
SF5	-0.147 (0.018)***	-0.217 (0.026)***	-0.195 (0.021)***	-0.195 (0.021)***	-0.419 (0.055)***	-0.200 (0.021)***	-0.520 (0.073)***	
PA2	-0.110 (0.021)***	-0.133 (0.028)***	-0.096 (0.024)***	-0.096 (0.024)***	-0.188 (0.050)***	-0.094 (0.024)***	-0.244 (0.058)***	
PA3	-0.162 (0.018)***	-0.184 (0.024)***	-0.160 (0.021)***	-0.160 (0.021)***	-0.273 (0.048)***	-0.168 (0.021)***	-0.368 (0.054)***	
PA4	-0.117 (0.020)***	-0.165 (0.026)***	-0.131 (0.023)***	-0.131 (0.023)***	-0.327 (0.053)***	-0.135 (0.023)***	-0.411 (0.066)***	
PA5	-0.236 (0.019)***	-0.326 (0.026)***	-0.239 (0.021)***	-0.239 (0.021)***	-0.558 (0.066)***	-0.250 (0.022)***	-0.700 (0.081)***	
PA6	-0.321 (0.023)***	-0.438 (0.034)***	-0.347 (0.026)***	-0.347 (0.026)***	-0.756 (0.084)***	-0.352 (0.027)***	-0.966 (0.111)***	
MH2	-0.063 (0.017)***	-0.092 (0.023)***	-0.063 (0.020)***	-0.063 (0.020)***	-0.140 (0.043)***	-0.064 (0.021)***	-0.256 (0.063)***	
MH3	-0.045 (0.017)**	-0.085 (0.023)***	-0.051 (0.020)**	-0.051 (0.020)**	-0.179 (0.047)***	-0.055 (0.020)***	-0.269 (0.065)***	
MH4	-0.190 (0.019)***	-0.266 (0.027)***	-0.220 (0.023)***	-0.220 (0.023)***	-0.464 (0.062)***	-0.222 (0.023)***	-0.654 (0.089)***	
MH5	-0.248 (0.017)***	-0.366 (0.026)***	-0.269 (0.020)***	-0.269 (0.020)***	-0.647 (0.075)***	-0.274 (0.020)***	-0.841 (0.106)***	
VI2	-0.091 (0.020)***	-0.124 (0.026)***	-0.069 (0.023)***	-0.069 (0.023)***	-0.188 (0.052)***	-0.069 (0.023)***	-0.221 (0.057)***	
VI3	-0.129 (0.018)***	-0.175 (0.023)***	-0.136 (0.020)***	-0.136 (0.020)***	-0.213 (0.041)***	-0.139 (0.020)***	-0.280 (0.054)***	

VI4	-0.170 (0.018)***	-0.214 (0.025)**	-0.176 (0.021)***	-0.343 (0.053)**	-0.175 (0.021)***	-0.450 (0.069)***
VI5	-0.288 (0.018)***	-0.388 (0.027)***	-0.322 (0.021)***	-0.675 (0.077)***	-0.329 (0.022)***	-0.870 (0.109)***
Duration x PF2	0.004 (0.001)***	0.008 (0.002)***	0.003 (0.001)**	0.015 (0.004)***	0.003 (0.001)*	0.019 (0.005)***
Duration x PF3	0.002 (0.001)*	0.005 (0.002)***	0.001 (0.001)	0.008 (0.004)**	0.001 (0.001)	0.009 (0.004)**
Duration x PF4	0.006 (0.001)***	0.010 (0.002)***	0.005 (0.001)***	0.018 (0.004)***	0.005 (0.002)***	0.017 (0.004)***
Duration x PF5	0.006 (0.001)***	0.010 (0.002)***	0.005 (0.002)***	0.018 (0.005)***	0.005 (0.002)***	0.022 (0.005)***
Duration x PF6	0.011 (0.001)***	0.016 (0.002)***	0.011 (0.002)***	0.026 (0.005)***	0.011 (0.002)***	0.029 (0.005)***
Duration x RL2	0.003 (0.001)***	0.002 (0.002)	0.004 (0.001)***	0.008 (0.003)***	0.004 (0.001)***	0.006 (0.003)**
Duration x RL3	-0.001 (0.001)	-0.004 (0.001)***	0.000 (0.001)	0.002 (0.003)	0.000 (0.001)	0.002 (0.003)**
Duration x RL4	0.002 (0.001)*	0.001 (0.001)	0.002 (0.001)*	0.005 (0.003)*	0.002 (0.001)	0.006 (0.003)**
Duration x SF2	0.008 (0.001)***	0.010 (0.002)***	0.010 (0.002)***	0.015 (0.003)***	0.010 (0.002)***	0.018 (0.004)***
Duration x SF3	0.005 (0.001)***	0.007 (0.002)***	0.005 (0.001)***	0.010 (0.003)***	0.004 (0.001)***	0.011 (0.003)***
Duration x SF4	0.008 (0.001)***	0.012 (0.002)***	0.009 (0.001)***	0.021 (0.004)***	0.010 (0.001)***	0.025 (0.004)***
Duration x SF5	0.007 (0.001)***	0.011 (0.002)***	0.010 (0.001)***	0.021 (0.003)***	0.010 (0.001)***	0.027 (0.004)***
Duration x PA2	0.006 (0.001)***	0.008 (0.002)***	0.005 (0.002)***	0.009 (0.003)***	0.005 (0.002)***	0.012 (0.004)***
Duration x PA3	0.008 (0.001)***	0.008 (0.002)***	0.008 (0.001)***	0.011 (0.003)***	0.008 (0.001)***	0.017 (0.003)***
Duration x PA4	0.004 (0.001)***	0.006 (0.002)***	0.005 (0.002)***	0.013 (0.003)***	0.005 (0.002)***	0.017 (0.004)***
Duration x PA5	0.010 (0.001)***	0.014 (0.002)***	0.010 (0.001)***	0.023 (0.003)***	0.010 (0.001)***	0.027 (0.004)***
Duration x PA6	0.016 (0.001)***	0.021 (0.002)***	0.016 (0.002)***	0.034 (0.004)***	0.016 (0.002)***	0.046 (0.006)***
Duration x MH2	0.003 (0.001)***	0.005 (0.001)***	0.003 (0.001)**	0.006 (0.003)**	0.003 (0.001)**	0.014 (0.004)***
Duration x MH3	0.002 (0.001)**	0.005 (0.001)***	0.003 (0.001)**	0.011 (0.003)***	0.003 (0.001)**	0.017 (0.004)***
Duration x MH4	0.009 (0.001)***	0.013 (0.002)***	0.010 (0.001)***	0.019 (0.003)***	0.010 (0.002)***	0.030 (0.005)***
Duration x MH5	0.011 (0.001)***	0.018 (0.002)***	0.012 (0.001)***	0.028 (0.004)***	0.012 (0.001)***	0.037 (0.005)***
Duration x VI2	0.006 (0.001)***	0.008 (0.002)***	0.004 (0.002)***	0.012 (0.003)***	0.004 (0.002)**	0.013 (0.004)***
Duration x VI3	0.007 (0.001)***	0.010 (0.002)***	0.007 (0.001)***	0.010 (0.003)***	0.007 (0.001)***	0.013 (0.004)***
Duration x VI4	0.007 (0.001)***	0.009 (0.002)***	0.007 (0.001)***	0.013 (0.003)***	0.007 (0.001)***	0.015 (0.004)***
Duration x VI5	0.015 (0.001)***	0.020 (0.002)***	0.017 (0.001)***	0.034 (0.004)***	0.017 (0.001)***	0.044 (0.006)***
τ		0.877 (0.044)		1.322 (0.087)		1.534 (0.084)
Standard						

Deviation [†]								
Duration			0.154 (0.009)***	0.365 (0.046)***	0.166 (0.010)***	0.434 (0.061)***		
PF6			0.044 (0.010)***	0.092 (0.016)***	0.067 (0.009)***	0.232 (0.034)***		
RL4			0.017 (0.017)	0.083 (0.017)***	0.039 (0.007)***	0.067 (0.012)***		
SF5			0.013 (0.017)	0.101 (0.020)***	0.018 (0.009)**	0.117 (0.019)***		
PA6			0.062 (0.010)***	0.050 (0.012)***	0.065 (0.010)***	0.143 (0.025)***		
MH5			0.072 (0.007)***	0.132 (0.019)***	0.081 (0.007)***	0.200 (0.028)***		
VI5			0.057 (0.008)***	0.145 (0.023)***	0.061 (0.008)***	0.227 (0.033)***		
γ				-0.054 (0.035)		-0.155 (0.055)***		
Log likelihood	-8639	-8432	-8409	-8204	-8376	-8159		
Degrees of freedom	52	53	59	61	80	82		
AIC	17383	16970	16936	16530	16912	16482		
BIC (n= individuals)	17635	17228	17223	16826	17301	17165		
BIC (n = observations)	17816	17412	17427	17037	17578	16881		

Statistical significance noted at the 1% level (****), the 5% level (**) and the 10% level (*)

[†] STATA reports standard deviations as both positive and negative, but notes that the sign is irrelevant. Therefore, the absolute values are presented here

[‡] Here, the standard deviations for each of the random variables were generated to allow some comparability with other models generating standard deviation figures. These were generated using either the STATA *mixlcv*, *sd* or the *gmilcov*, *sd* command. The variance-covariance matrices for Models D5 and D6 are provided in Table 32 and Table 33 respectively.

In model D1, the pattern of coefficients relating to the non-linearity of the utility function is interesting. The quadratic term on duration is negative and statistically significant suggesting the utility associated additional years of life independent of quality is diminishing.

Additionally, the coefficients on the interactions between duration and the levels of the SF-6D are generally positive and statistically significant. The interpretation of these is that, for longer durations, quality of life becomes relatively less important.

Model D2 identifies that, as with Model A2 relative to A1, introducing a scale parameter has a significant benefit in terms of model fit. The one additional parameter improves the log-likelihood by over 200; this compares with the mixed logit in model D3 which improves the log-likelihood by 230 points relative to the conditional logit but with an additional 7 degrees of freedom. In this non-linear setting, the support for the mixed logit relative to the S-MNL is equivocal; AIC improves, but BIC does not.

Combining scale and taste heterogeneity in a model without correlation (i.e. model D4) appears to improve on either D2 or D3 as determined by the Information Criteria. The inclusion of scale in a model with taste heterogeneity already considered (i.e. moving from A3 to A4) has a large impact on log-likelihood, reiterating the importance of allowing for scale in this context. It should also be noted that the value for γ is low suggesting that the scale term applies almost equally to the parameter coefficient β and the variance term η_i (Fiebig's (2010) G-MNL-II). Introducing correlations to Model D3 (i.e. with Model D5) improves log-likelihood by 33, but with an additional 21 degrees of freedom. The AIC suggests the additional model fit is worthwhile, but the BIC contradicts this. Model D6 is superior to D5, but, as has been observed already, the value of allowing a small set of correlated coefficients is uncertain (in that the BIC is better under the more restricted D4 than D6, while the AIC suggests D6 is superior).

To summarise, as with Models A1-A6, the introduction of scale and preference heterogeneity appears to generally improve the model fit. However, the benefit of allowing correlated coefficients beyond these is uncertain, given that Model B3 and B4 appear to outperform Models B5 and B6 respectively under the Bayesian Information Criteria.

Relative to utility function A, utility function D does best in the more restricted models such as 1. In B1, the extra explanatory capability of the regression justifies the additional parameters in AIC and BIC terms. This pattern is repeated across models 2-6, but the

absolute difference in Information Criteria decreases as the model becomes increasingly less constrained.

The variance-covariance matrices for Models B5 and B6 are provided in Table 32 and Table 33.

Table 32: Variance-Covariance Matrix for Model B5

	Duration	PF6	RL4	SF5	PA6	MH5	VI5
Duration	0.166 (0.010) ***						
PF6	-0.026 (0.007) ***	0.062 (0.009) ***					
RL4	-0.012 (0.005) **	0.033 (0.008) ***	-0.019 (0.007) ***				
SF5	0.006 (0.006)	-0.012 (0.009)	-0.005 (0.008)	-0.011 (0.009)			
PA6	-0.018 (0.008) **	0.014 (0.011)	0.037 (0.011) ***	-0.040 (0.012) ***	0.027 (0.015) *		
MH5	-0.011 (0.006) *	0.013 (0.012)	-0.066 (0.008) ***	-0.033 (0.019) *	0.004 (0.013)	0.028 (0.023)	
VI5	-0.002 (0.007)	0.022 (0.008) ***	0.004 (0.010)	-0.004 (0.013)	-0.019 (0.012)	0.028 (0.012) **	-0.045 (0.01) ***

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

Standard errors noted in parentheses

Table 33: Variance-Covariance Matrix for Model B6

	Duration	PF6	RL4	SF5	PA6	MH5	VI5
Duration	0.434 (0.061) ***						
PF6	-0.047 (0.013) ***	0.227 (0.033) ***					
RL4	0.016 (0.008) **	0.064 (0.011) ***	-0.012 (0.009) **				
SF5	0.024 (0.011) **	-0.02 (0.01) **	0.082 (0.015) ***	-0.078 (0.014) ***			
PA6	0.016 (0.013) **	0.008 (0.013) **	0.029 (0.015) **	0.019 (0.015) **	0.137 (0.023) ***		
MH5	0.072 (0.014) ***	0.013 (0.013) **	0.013 (0.01) **	0.022 (0.01) **	-0.107 (0.017) ***	-0.151 (0.022) ***	
VI5	-0.017 (0.012) **	0.097 (0.016) ***	0.077 (0.015) ***	0.156 (0.024) ***	-0.093 (0.016) ***	0.039 (0.013) ***	0.039 (0.012) ***

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

Standard errors noted in parentheses

Overall model comparisons

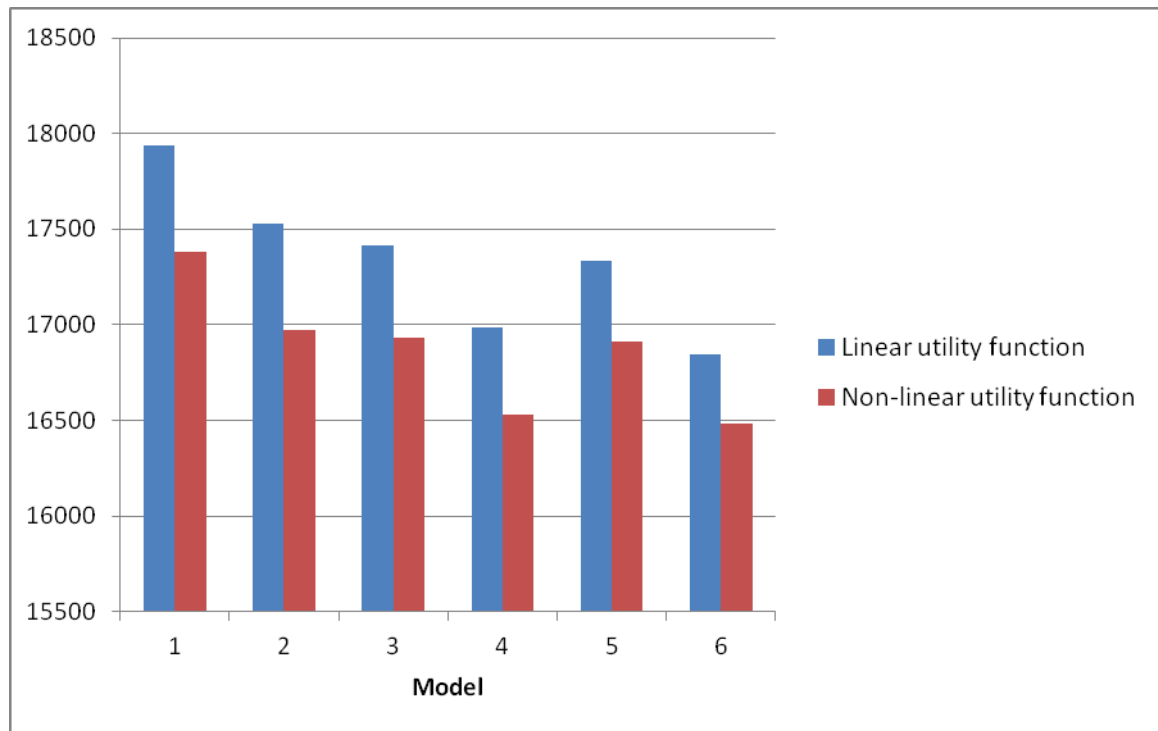
The summary of AIC and BIC under each of the twelve regressions are presented in Table 34.

Table 34: Model Comparison

Model	AIC	BIC (n=individuals)	BIC (n=observations)
A1	17938	18066	18136
A2	17530	17530	17755
A3	17415	17574	17598
A4	16984	17069	17193
A5	17337	17598	17787
A6	16849	17120	17315
B1	17383	17635	17816
B2	16970	17228	17412
B3	16936	17223	17427
B4	16530	16826	17037
B5	16912	17301	17578
B6	16484	16883	17167

The AIC figures are displayed graphically in Figure 23.

Figure 23: Comparison of Akaike Information Criteria (AIC)



The BIC figures where n =observations or individuals are displayed graphically in Figure 24 and Figure 25.

Figure 24: Comparison of Bayesian Information Criteria (BIC) (n =observations)

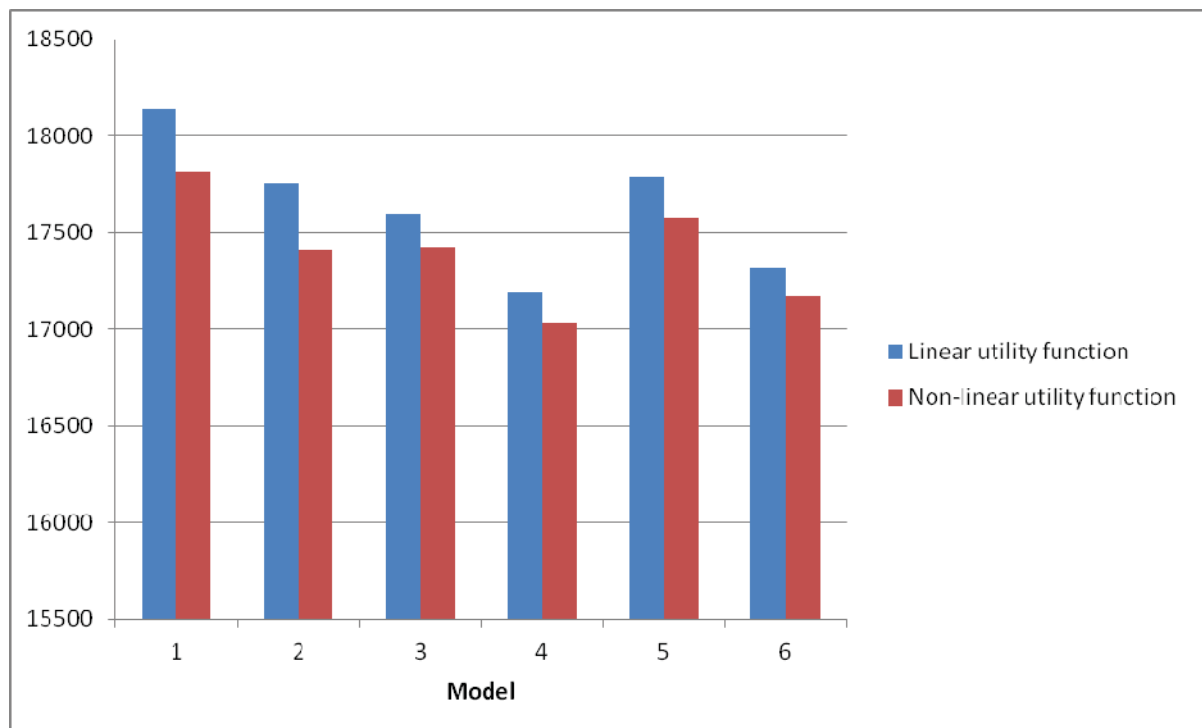
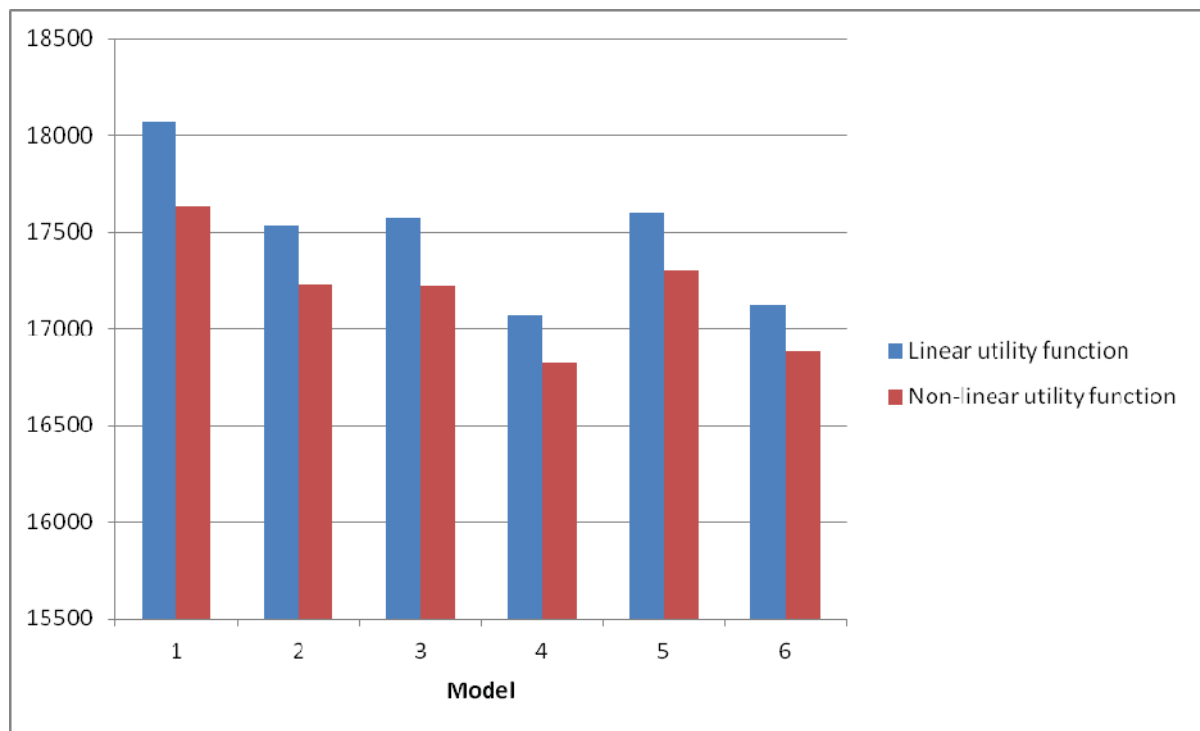


Figure 25: Comparison of Bayesian Information Criteria (BIC) (n=individuals)



There are four broad conclusions in these data. Firstly, in all models, the non-linear Utility Function D outperforms Utility Function A (which imposes the QALY model on the data). However, as the approach to modelling heterogeneity becomes more sophisticated, the gap between the BIC under Utility Functions A and D becomes smaller. This conclusion relates simply to the model fit; the problems associated with a more complicated QALY algorithm have to be addressed before these weights could be used in economic evaluation.

Secondly, applying some consideration of the panel nature of the data is beneficial in model fit terms, as noted by the consistent improvement in BIC under models 2 and 3 (and indeed, 4-6) relative to Model 1.

Thirdly, the models which add consideration of a scale effect (2, 4 and 6) have large gains in BIC relative to the model without that consideration (1, 3 and 5 respectively). This is because they both increase model fit, and add a relatively small number of additional degrees of freedom (τ in 2, 4 and 6, and γ in 4 and 6 only).

Fourthly and finally, the allowance of some correlation between coefficients is of questionable value. Permitting correlations between a subset of the worst levels of each dimension of the SF-6D (i.e. moving from Model 3 to 5, or from 4 to 6) has a deleterious effect on BIC. This pattern is opposite to that seen in the AIC, but given the difficulty in

selecting and modelling these coefficients, it is likely that the uncorrelated case is adequate in this circumstance.

Deriving utility weights under models A1-A6

At the start of the section, two distinct tasks were outlined. These were the derivation of population weights for generic quality of life states and the modelling of respondent heterogeneity. In identifying a preferred model, there is a tension between the two. In terms of fitting the model to the data, a more relaxed approach is recommended. Allowing a non-linear utility function with respect to time is justified, as is adjusting for both scale and preference heterogeneity. However, a non-linear utility function means QALY weights for generic quality of life instruments are dependent on the period of time under consideration (which is atypical and introduces complexity into economic evaluation), and the inclusion of heterogeneity has little impact on the population mean response. This relates to the conclusion of Greene and Hensher (2011), who conclude that, while accounting for scale impacts on model fit, it has little impact on estimates of welfare measures, for example willingness to pay. The QALY weights generated by all models are given in Table 35.

Table 35: DCE-derived QALY Weights for the SF-6D (Main Effects Only)

Model	A1	A2	A3	A4	A5	A6
PF2	-0.042	0.011	-0.035	-0.019	-0.034	-0.015
PF3	-0.078	-0.068	-0.071	-0.058	-0.058	-0.056
PF4	-0.135	-0.117	-0.128	-0.116	-0.126	-0.118
PF5	-0.141	-0.116	-0.137	-0.120	-0.134	-0.119
PF6	-0.288	-0.241	-0.276	-0.243	-0.285	-0.250
RL2	-0.096	-0.089	-0.086	-0.065	-0.076	-0.081
RL3	-0.069	-0.059	-0.055	-0.035	-0.044	-0.037
RL4	-0.126	-0.133	-0.113	-0.101	-0.114	-0.110
SF2	-0.036	-0.076	-0.029	-0.038	-0.024	-0.036
SF3	-0.039	-0.063	-0.042	-0.053	-0.042	-0.048
SF4	-0.117	-0.123	-0.110	-0.101	-0.104	-0.097
SF5	-0.129	-0.148	-0.135	-0.125	-0.126	-0.138
PA2	-0.078	-0.064	-0.091	-0.059	-0.092	-0.084
PA3	-0.168	-0.163	-0.157	-0.125	-0.152	-0.156
PA4	-0.198	-0.189	-0.185	-0.138	-0.182	-0.172
PA5	-0.288	-0.301	-0.276	-0.235	-0.275	-0.274
PA6	-0.270	-0.265	-0.272	-0.252	-0.287	-0.288
MH2	-0.063	-0.051	-0.049	-0.062	-0.044	-0.066
MH3	-0.069	-0.030	-0.038	-0.035	-0.036	-0.022
MH4	-0.180	-0.165	-0.177	-0.173	-0.182	-0.193
MH5	-0.267	-0.242	-0.249	-0.229	-0.261	-0.253
VI2	-0.009	-0.004	-0.009	-0.010	-0.014	-0.008
VI3	-0.045	-0.040	-0.044	-0.042	-0.050	-0.049
VI4	-0.213	-0.201	-0.179	-0.152	-0.182	-0.172
VI5	-0.252	-0.229	-0.230	-0.197	-0.232	-0.230

As with the random-effect probit results, there are a number of coefficients on levels which appear to demonstrate a non-monotonic ordering. As the purpose of this analysis was more to consider the agreement between mean results under different approaches to modelling heterogeneity, no pooling was undertaken on these results.

There is considerable agreement between the QALY weights associated with the estimated models. If the 18,000 SF-6D health states are valued using each algorithm, the correlation coefficients and Spearman rank coefficients are given in Table 36 and Table 37.

Table 36: Correlation Coefficients for the 18,000 Health State Valuations

	A1	A2	A3	A4	A5	A6
A1	1.000					
A2	0.990	1.000				
A3	0.996	0.988	1.000			
A4	0.987	0.987	0.995	1.000		
A5	0.992	0.984	0.998	0.996	1.000	
A6	0.986	0.987	0.995	0.997	0.995	1.000

Table 37: Spearman Rank Coefficients for the 18,000 Health State Valuations

	A1	A2	A3	A4	A5	A6
A1	1.000					
A2	0.989	1.000				
A3	0.995	0.988	1.000			
A4	0.986	0.985	0.994	1.000		
A5	0.991	0.983	0.998	0.996	1.000	
A6	0.984	0.986	0.994	0.997	0.995	1.000

As shown in Table 36 and Table 37, the utility weights associated with each of the six models are almost perfectly correlated. While the more advanced specifications accounting for heterogeneity lead to a considerable improvement in terms of model fit, employing them does not affect inferences regarding the scores placed on individual health states. Thus, while it is interesting to note the heterogeneous results provided by respondents, the impact of doing so on the mean respondent is minimal.

Chapter discussion

In this chapter, it has been shown that the construction of QALY weights using a DCE with efficient designed experiment properties is feasible and produces results which both predict choices well and avoid some of the criticisms which can apply to other preference elicitation tasks such as the Time Trade Off and Standard Gamble. It would be valuable to use this approach in a larger sample, and with a design which allows for unbiased estimation of higher-order interaction terms (most likely three-factor interactions including duration).

One issue which needs addressing is whether the weights generated comply with the constraints required to be defined as QALY weights. Bleichrodt *et al.* identified that QALYs require only one contentious condition, that being risk neutrality (Bleichrodt, et al., 1997). They showed that risk neutrality implies both the zero condition and constant proportional

trade-offs which have generally been considered part of the characterisation of the QALY model (Pliskin, et al., 1980). Constant proportional trade-offs were not shown in the logistic regression as the quadratic term on duration was significant and negative suggesting diminishing importance of extra life expectancy. However, the approach taken here i.e. relaxing the assumption of constant proportional trade-offs, increases the predictive value of the model. In addition, it is not likely that somehow imposing the constraints on these weights would affect the trade-off between dimensions of the SF-6D.

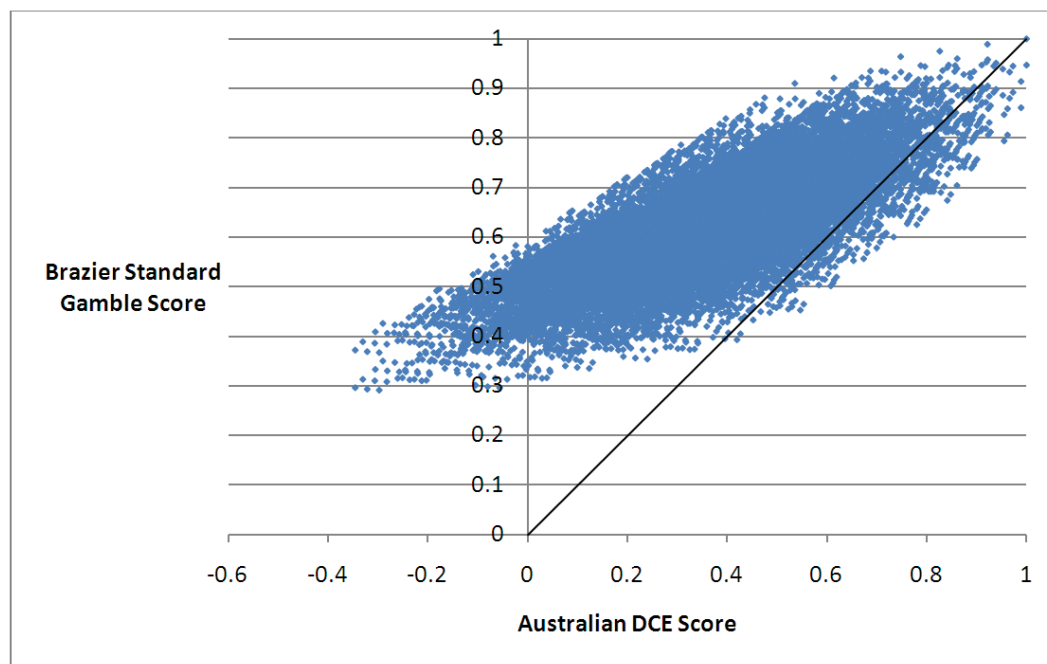
One potential limitation of the approach adopted in this study is the possibility that an online panel is not representative of the Australian population, which may limit the applicability of the weights to the Australian population overall. An online panel was used because it is a very cost-effective means of collecting these data, and these approaches are increasingly used in valuation studies (Wittenberg and Prosser, 2011). While the panel respondents can be selected to enhance the representativeness on observable characteristics, there is still the potential concern that these respondents differ from the general population in some unadjusted or unobservable dimension. However, we would argue that this criticism applies to some degree to all approaches to survey administration that might be used. The second criticism is that it has been argued that the mode of administration can significantly affect the quality of data collected (Bowling, 2005). Potentially, the nature of online respondents make this approach particularly susceptible to poor data quality in that they are not being observed while responding and it is not possible to identify how carefully they are considering the choices. However, it is likely that the weight of this more general criticism depends on the nature of the task. In a DCE, answering on criteria above and beyond the levels and dimensions presented (at extreme this might be answering randomly, or answering all A's for example) does not bias the results assuming that some basic design properties have been considered. Thus, it is important to limit any unconsidered responses, but the impact of these in a DCE is to reduce the effective sample size rather than systematically bias the conclusions drawn. This contrasts with the TTO in which unconsidered responses do systematically bias conclusions, an example of which has been described elsewhere (Norman, et al., 2010).

Finally, it should be noted that the weights derived here differ in some respects from those in the original Brazier valuation study (2002), and also in the non-parametric approach of Kharroubi (2007). In the DCE approach presented here, 12% of health states described by the SF-6D were valued below zero (meaning they are considered to be worse than death).

This contrasts with the floor effect observed by Brazier *et al.* (2002) in which no health states are valued as being worse than death, and indeed the minimum value is approximately 0.3.

A scatter plot contrasting the utility weights placed on the 18,000 SF-6D states under Brazier's UK algorithm and the Australian DCE algorithm developed in this work is presented in Figure 26, with a black line of equality drawn to allow comparison.

Figure 26: Comparison of Health State Valuation under Different Algorithms



There is clearly a degree of agreement between the two algorithms (and hence in the valuations placed on individual states). However, it is also apparent that scores under Brazier's algorithm are above those assigned using the DCE methods described here, a trend which is increasingly strong for poorer health states.

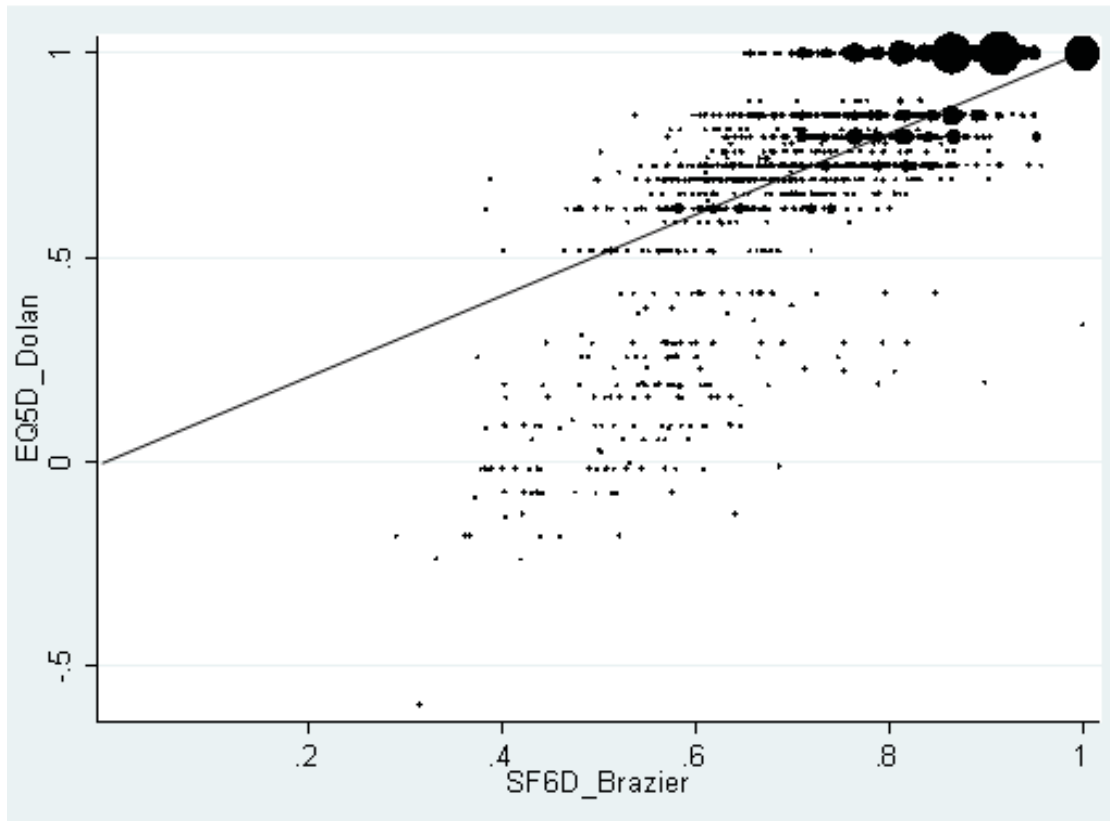
The importance of any divergence is likely to lie in the use of such weights in economic evaluation. The range of weights constructed using this DCE is higher than those generated by Brazier's Standard Gamble, giving a broader scope for a healthcare intervention to improve quality of life, and an increased gain in a generic outcome metric such as a QALY. This means that, on average, using DCE weights gives larger gains for any given improvement in quality of life as described by the SF-6D, with implications for the cost per QALY. As there is not yet a definitive approach to quality of life measurement and valuation. In England and Wales, the National Institute for Health and Clinical Excellence

(NICE) has proposed a reference case, based on the EQ-5D instrument and the algorithm developed by Dolan (1997). The motivation behind selecting a single instrument and a single algorithm is to allow comparability across interventions. Clearly, as raised in Chapter 2, there are issues relating to the sensitivity of instruments to particular types of change in quality of life (Hawthorne, et al., 2001). In terms of selection of instrument, there may be issues with comparability. In terms of choice of algorithm, it is plausible that recommending a common algorithm be used will promote comparability. However, the question remains whether any algorithm can be determined to be the most appropriate of those available. There are a variety of possible approaches to take, each with particular strengths and weaknesses. In terms of implications for reimbursement decisions, my recommendation would be that cost-effectiveness conclusions be tested under the range of available algorithms; the kind of study presented here adds to the suite of potential valuation sets which might be included in this robustness testing.

The final question posed in this chapter is “To what extent the use of a common method for eliciting preferences (such as a DCE) increases convergence in self-assessed health scores between different generic multi-attribute utility instruments?”. To state the issue differently, it has been shown in Chapter 2 that self-assessed health is valued very differently between the EQ-5D (with preferences measured using the TTO) and the SF-6D (with preferences measured using the Standard Gamble). If the method of valuation is standardised using a third technique (the DCE), what divergence, which would best be explained by differences in the instrument itself, remains? Whitehurst and Bryan (2011) argue that differences in weights associated with self-assessed health between the EQ-5D and SF-6D are unlikely to be removed even if a common preference elicitation technique is applied. This echoes the argument made by Konerding *et al.* (2009), who claim that the EQ-5D and SF-6D would “*produce different valuations even if these valuations were determined according to the same principle*” (p.1249). While I agree that differences will inevitably remain due to the selection of dimensions within each instrument and the wording of specific levels, it is highly plausible that removing one of the differences between the valuation techniques for the two instruments would bring the scores assigned to a particular person under the two instruments closer together. The self-assessed health described in Chapter 2 is therefore combined with the SF-6D algorithm described here, and the EQ-5D algorithm described by Viney *et al.* (2011a).

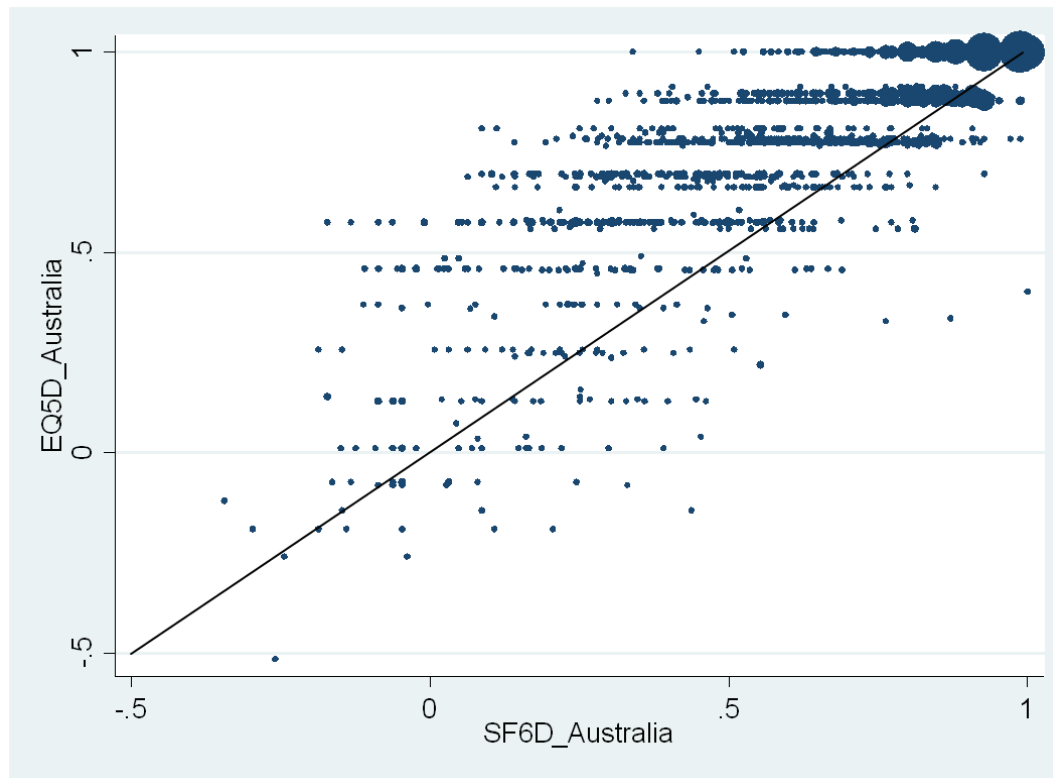
Figure 2 in Chapter 2 contrasted the utility weights assigned to individuals using the SF-6D algorithm of Brazier (2002), and the EQ-5D algorithm of Dolan (1997). This is reproduced here as Figure 27.

Figure 27: Comparison of utility weights associated with general population sample using pre-existing SF-6D and EQ-5D algorithms



The conclusion to be made from these data was that the weight assigned to the health of an individual was strongly determined by the choice of MAUI (with the selection of valuation technique associated with each). The results of this analysis are now repeated for the two DCE-derived sets of weights, and are presented in Figure 28.

Figure 28: Comparison of utility weights associated with general population sample using Australian DCE-derived algorithms



In those patients reporting no problems in the EQ-5D, but some problems in the SF-6D, the difference in utility weight is smaller using the DCE weights. This can be seen as the clusters of individuals at the highest point of the y-axis are generally closer to the (1,1) point in Figure 28 than in Figure 27.

The tendency for the EQ-5D to have a broader range of values that was observed in Chapter 2 (in which the TTO was used to derive valuations for health states) has largely been eliminated by adopting a common valuation technique. A summary of the agreement between EQ-5D and SF-6D utility weights using either the original UK algorithms or the Australian DCE-derived ones are presented in Table 38.

Table 38: Agreement between instruments under existing and novel methods

	Existing methods	Australian DCE methods
SF-6D score =	0.4450+0.4066(EQ5D)	-0.1049+0.9650(EQ5D)
R ²	0.5767	0.6682
Adjusted R ²	0.5765	0.6680
Spearman coefficient	0.7413	0.7929
Correlation coefficient	0.7594	0.8174
Mean difference (SF-6D – EQ-5D)	-0.0311 (SD: 0.1618)	-0.1339
Mean absolute difference	0.1193 (SD: 0.1136)	0.1534 (SD: 0.1342)
Mean squared difference	0.0271 (SD: 0.0588)	0.0415 (SD: 0.0663)

Under the DCE methods presented here, the simple OLS regression presented in Table 38 suggests the value placed on self-assessed health under the EQ-5D and the SF-6D has an almost 1-to-1 relationship (in that the constant is small and the coefficient on the EQ-5D weight is close to 1). This evidence in favour of the novel DCE methodology is supported by improved R² values and Spearman and correlation coefficients. However, the evidence regarding mean differences (raw, absolute or squared) is less favourable for the DCE methods. This reflects the trend in Figure 28 for the EQ-5D weight to generally lie above that of the SF-6D.

Chapter 6: Equity Weights for Use in Economic Evaluation

Chapter summary

In this chapter, I present an investigation of an area in which conventional valuation of changing health outcomes may differ from true societal preferences. This relates to the different values placed on health outcomes depending on to whom they accrue, an area which, while it does play a role in decision making currently, is distinct from the health maximisation framework implied in conventional economic evaluation. The chapter begins by considering the role that equity currently plays in decision making. The concept of equity weights will then be introduced, and given a theoretical framework. Then, a discrete choice experiment will be run investigating the kinds of trade-offs that people are willing to make between health improvements in different groups. This will investigate different approaches to modelling heterogeneity of responses, and contrast them in terms of model fit and parsimony.

Introduction to equity in economic evaluations of health care interventions

In Chapter 1, the conventional concept of economic evaluation in healthcare was discussed. It was identified that the most conventional aim was to maximise the total health of the population, rather than to the maximisation of utility which is the aim of economic evaluation in a welfarist framework. Under either a welfarist or an extra-welfarist framework, any equity considerations were placed alongside (but distinct from) economic considerations. This basis of economic evaluation does not recognise that there may be considerable differences in the value society places on health gains dependent on to whom they accrue. It should be noted that, as the outcome is in terms of health alone, there is no scope for traditional Paretian compensation of losers by gainers (Coast, 2009).

Sassi *et al.* (2001). argue not just that the consideration of equity as part of economic evaluation is limited, but that making normative judgements as part of economic evaluation of healthcare posed “*significant, if not insurmountable, theoretical and practical problems*”(p.7).

If this is the case, one obvious solution is to present cost-effectiveness evidence alongside other pertinent information (relating to safety, budgetary implications, distributional impact etc), and allow decision makers to informally balance these against each another. This is the approach which is adopted in decision making processes worldwide. However, this is

arguably somewhat unsatisfactory as the method for balancing between these areas is unclear.

A number of competing concepts of equity in healthcare have been considered. Sen (1980; 1992) argues that normative theories of social distribution more generally are all based on some concept of egalitarianism. If equity in healthcare is based on egalitarianism, the question is then in which domain egalitarianism should be measured. There are a number of competing forms of egalitarianism that have been suggested, including equality of outcome (be it life expectancy, quality-adjusted life expectancy or some other measure), equality of gain in outcome, equality of access, equality of resource or equality of opportunity. It is clear that equality in one of these dimensions does not imply equality in another. Therefore, it follows that an advocate of one form of egalitarianism in health and health care must be willing to accept inequalities in a different dimension. Hausman and McPherson (1996) summarise this point,

“What makes moral theories so different is that the things different moral theorists seek to equalize are not perfectly correlated with one another. Equalizing one thing conflicts with equalizing another” (p.135)

Thus, from any egalitarian viewpoint, there are inequalities considered equitable, or at least acceptable. This point reflects Dworkin’s work (1977) that asserted that conflict between political theories can best be understood as conflicting interpretations of equal respect.

In this chapter, the emphasis will be on equity as equality of health outcome. As discussed in the introductory chapter, this focus on equality of outcome is somewhat contentious as it ignores issues such as access and capabilities which are the crux of certain approaches to considering healthcare decision making (see for example Sen (1980) or Mooney (1991)). Additionally, it largely ignores the value of information provision, and consequently issues such as reassurance. Nevertheless, to explore equity issues in a quantitative manner requires a view of what equity consists of to be stated.

Equity and altruism

Before looking at existing approaches to quantifying the efficiency-equity trade-off, it is necessary to define exactly what is meant by equity in this chapter. Wagstaff and van Doorslaer (2000) make the helpful distinction between equity and altruism, arguing that the two concepts are often (erroneously) used interchangeably. Altruism is a matter of

preference; a person is altruistic if they are willing to forego some of their resources to improve the outcome for another person. While altruism (or caring) is seemingly very different from conventional concerns in economic evaluation, Culyer (1989) makes the point that it fits neatly within the language of efficiency.

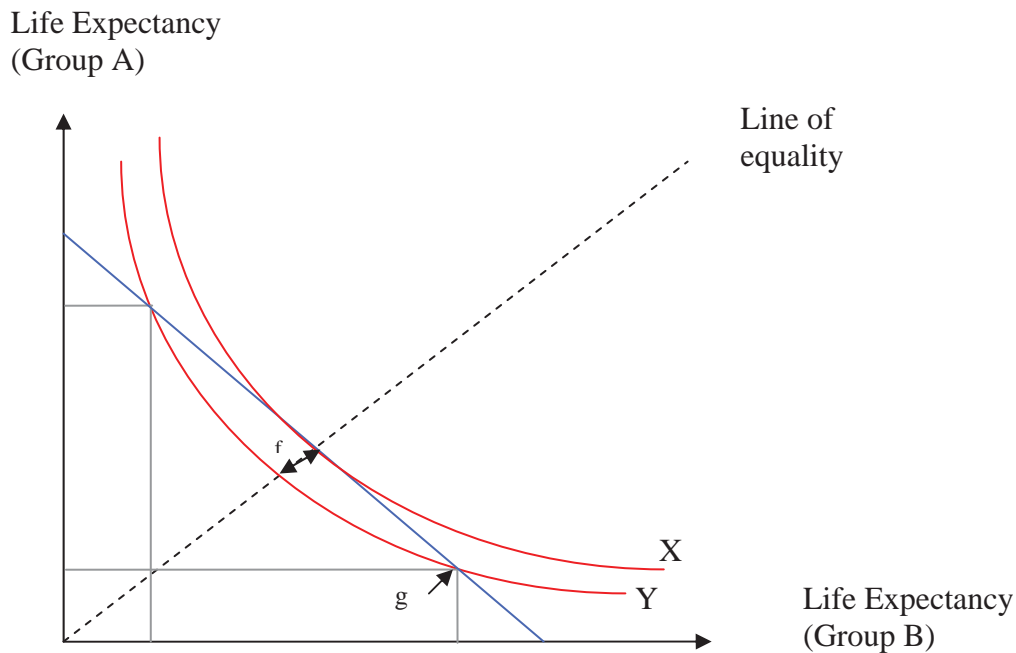
In contrast, equity exists outside of preferences, and reflects what a person ought to have by right. Correctly, Wagstaff and van Doorslaer identify this is difficult to do within a standard survey as self-interest impacts on how people respond. They suggest that a Rawlsian ‘veil of ignorance’ approach might help (Rawls, 1999). In this, the respondent is placed in a position of not knowing which position they will inhabit in a society, and then asked about appropriateness of the distribution of resources within this society. To identify this as a helpful approach in certain circumstances is correct; when looking at a distribution of some outcome across a population of anonymous individuals, this approach is tractable and helpful. However, as will become clear in the description of asymmetric preferences below, there are instances where such an approach is not possible.

Social Welfare Functions and equity

Economic evaluation of health technologies and interventions has been conventionally based on the assumption that a gain is of equal value irrespective of to whom it accrues. This is a standard assumption within either the Life Year or Quality-Adjusted Life Year (QALY) model. Among the sequelae of the utilitarianism approach implicit in both of these models, two important points need to be made with regard to its implication for a social welfare function (SWF)¹⁰. Firstly, it assumes that the SWF between the health of two groups of equal size is linear with a slope of -1. That is, a unit of outcome for one individual is a perfect substitute for a unit of outcome for another individual. This means that, from the perspective of the economist, the distribution of outcomes is unimportant. A typical SWF under a utilitarian approach is given in Figure 29 as the blue line.

¹⁰ The use of the term Social Welfare Function does not preclude the importance of other issues beyond the life expectancy of the two groups in the aggregation of social welfare. It is assumed that it does represent the SWF and that all other considerations are held constant.

Figure 29: Symmetrical Utilitarian and non-Utilitarian Social Welfare Functions



Under the blue SWF, points f and g are of equal value as they have the same total health, even though g implies considerable inequality of life expectancy. If linearity is considered inappropriate (meaning distribution matters), a class of SWF such as those given in red are one possible alternative. Introducing concavity into the utility function imposes a degree of inequality-aversion reflected by the severity of the curvature. An aversion to inequality in life expectancy would lead to the relative value of a year of life for person A being dependent on the life expectancy of both person A and person B. The implication of the class of SWF typified by X and Y is clear. While SWF X meets the blue life expectancy-maximising SWF at the point of equality, any move away from equality along the SWF would move to a higher life expectancy maximising SWF (or conversely, any movement along the blue SWF from the point of equality meets a SWF worse than X). Therefore, under X and Y, society is willing to sacrifice total life expectancy to ensure a more even distribution. Under Y, to move from an unequal point (marked g) to the line of inequality, society is willing to sacrifice f . Thus, increased curvature of the SWF represents an increased aversion to inequality (as increased curvature of the SWF passing through g increases the size of f). In the extreme, an L-shaped SWF describes an extremely inequality averse society in which the life expectancy of A and B are perfect complements¹¹.

¹¹ A more inequality-averse SWF is possible in which additional health to a group with initially better health might be viewed as a bad (rather than an irrelevance as would be the case under an L-shaped SWF).

Criticisms of SWF linearity

Criticism of the QALY (or life-expectancy) maximising model (as typified by the blue line) has come in a number of forms. With regard to the assumption of linearity, authors have argued that there are a number of reasons why linearity is either ethically indefensible, unrepresentative of actual societal preferences, or both. Regarding how representative a linear SWF is of public preferences, Dolan *et al.* (2005) reviewed the literature base regarding linearity in both quantity and quality of life and identified a clear and persuasive consensus that both decline in marginal value as they increase. However, evidence that society does not respond in a particular way (i.e. maximising total life expectancy or QALYs) does not immediately necessitate a different approach to societal decision making. Firstly, if society is to apply the preferences revealed by the majority, it is arguable that these preferences must be morally defensible (Richardson, 2002a). As Olsen *et al.* (2003) note, this is in keeping with Broome's idea of using laundered preferences (1991). Broome argues that the view of the society must be morally defensible from some *a priori* position if it is to become policy. However, this view might be countered by arguing that the acceptance of a view as morally defensible is difficult to disentangle from the general support for it. Thus, the view of the majority will rarely be laundered out, particularly in a situation where the decision maker is elected by society (and therefore ought, or is likely, to act to a large degree as an agent of their constituents). Additionally, if there is to be a system of laundering applied to average population preferences, the question remains regarding who does this. A sole arbiter of moral acceptability is precariously close to dictatorship which is not a clearly superior solution to accepting the view of the majority of the population however distasteful.

Tsuchiya (2000) provides a number of justifications for discriminating in the context of a symmetrical Social Welfare Function. Of these, the two most convincing are Daniels' Prudential Lifetime Account (1988) and Williams' Extended Fair Innings Argument (1997). Daniels argues that the question of resource allocation between the old and the young should be reframed as an allocation of resources over the lifetime of an individual. As Tsuchiya notes (while outlining Daniels' view),

“The purpose of health care is to secure a fair equality of opportunity for everybody, and this implies that resources ought to be allocated so that each can achieve a ‘normal lifespan’... prudential deliberators will choose to give priority to as many people as possible in order to allow them to reach the normal lifespan” (pp.60-61)

Williams (1997) frames his argument, which leads to similar conclusions, in terms of fairness rather than prudence, and stems from the basic notion that “*Death at 25 is viewed very differently from death at 85*” (p.119). Williams considers the possibility of applying a weight of greater (less) than 1 to an individual if their expected lifetime QALY at present age is below (above) a fair innings. There is good evidence for the Fair Innings argument playing a central role in the preferences of individuals for healthcare decision making, ahead of other concepts of equity such as proportional shortfall or severity of illness (Stolk, et al., 2005). The fair innings is in turn defined as the point of the SWF where lifetime expected QALYs are the same for all individuals. This definition of fairness for the fair innings is problematic if expected lifetime QALYs are a function not only of personal characteristics which are not chosen, but also of those which the individual makes an active decision to adopt, such as the decision to participate in dangerous activities, or not to take adequate stewardship of their own body. Of course, this objection is dependent on being able to identify those characteristics that are the choice of the individual, and those which are not.

A second reason why a move away from the simple SWF implicit in the blue line in Figure 29 is the practicality of explicitly including a non-linear SWF in decision making. While it may better represent societal preferences, the difficulties around estimating it may produce considerable uncertainty in its appropriateness. This is the major concern of this chapter and will be covered in depth later.

Symmetry of the SWF

The second assumption embedded within the standard economic evaluation approach is that the SWF, linear or otherwise is symmetrical around a forty-five degree line from the origin. Therefore, the only consideration in discriminating between groups is any difference in outcome (or expected outcome). Even if we allow a non-linear SWF, most of the previous attempts to quantify the efficiency-equity trade-off have focused on symmetrical preferences (Bleichrodt, et al., 2004; Williams, 1997). However, as noted previously, stated preference surveys have shown some tentative patterns of responses identifying that society is willing to weight outcomes based on characteristics beyond life expectancy (Olsen, et al., 2003). If a system of weights which represents this asymmetric preference set can be developed, it may (depending on the asymmetry of preferences) represent a significant improvement relative to either the QALY maximising or the non-linear symmetrical model. This argument is illustrated in Figure 30.

Figure 30: Relaxing the symmetrical assumption in non-linear SWF's

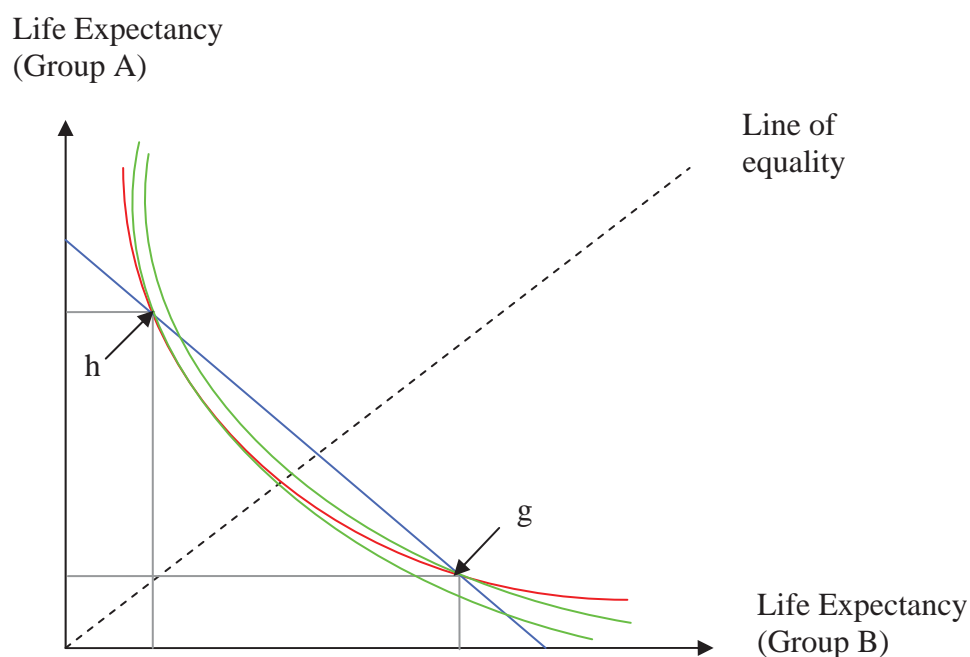


Figure 30 has the same life expectancy maximising SWF in blue and symmetrical (but non-linear) SWF in red that was used in Figure 29. Figure 30 contrasts this with a SWF in which one group or individual is valued more highly than another (in green). Olsen *et al.* (2003) identify some groups which may be relatively favoured in a representative SWF, including people with children, the employed and the poor. Under the blue SWF, points g and h are identical as the sum of the life expectancy is the same. As identified in Figure 29, the symmetrical non-linear SWF is willing to sacrifice some aggregate life expectancy to gain more equality. For the society represented by the red SWF, points g and h are equally valued as the gap f (as shown in Figure 29) is constant. However, this finding does not hold if we allow non-symmetrical SWF's (in green). The life expectancy of group B is for some reason, valued more highly than that of group A. Point h is now on a lower SWF than point g , meaning that the inequality at point h is more unpalatable than that at point g . This can be expressed by considering the amount that society would be willing to sacrifice to remove all inequality (gap f in the symmetrical preferences example). When the less valued group are disadvantaged in terms of life expectancy (point g), the gap between the intersection of the green SWF and the blue life expectancy maximising SWF is smaller than in the symmetrical

case. Equally, if the lower life expectancy is in the more favoured group (point *h*), the willingness to sacrifice total life expectancy to achieve equality is larger.¹²

The aim of this chapter is therefore:

- To evaluate the existing literature regarding the preferences of the general population to discriminate between groups and individuals
- To consider the viability of generating equity weights while relaxing the assumptions of linear and symmetrical SWFs. This includes questions of how best to survey the general population, and how to analyse their responses
- To consider modelling approaches, particularly relating to observable and unobservable heterogeneity
- To discuss whether the revealed preferences can be justified ethically (in Broome's terms, whether they reflect laundered preferences or not)
- To judge whether this approach can be integrated into economic evaluation of health interventions and approaches, and how such an integration might be done

Identifying relevant literature

A literature search was undertaken using Medline and Embase through the Ovid interface. The search was undertaken in May 2009. The purpose of the literature search was two-fold. The first aim was to identify papers that investigate the principle of trading off between different individual characteristics when making healthcare policy decisions. These characteristics are those which would not be considered under the QALY-type approach (so do not for example include issues such as an individual's capacity to benefit from an intervention, or any characteristic which may influence the cost that accrues as a result of the provision of the intervention). The second aim of the search were studies which discussed a specific trade-off, for example one which considered the relative worth of an additional year of life for a smoker and a non-smoker.

¹² It should be noted that there is a case in which inequality is neither actively viewed negatively nor irrelevant, but as a positive. The SWF in this case would be convex. However, this would only occur if society was inequality loving in general, or that they valued a group so lowly that increasing outcomes to that group was actually viewed as a negative. However, neither of these is likely (and are also likely to be laundered out if Broome's approach is taken).

The construction of the search strategy posed significant obstacles. The first step was to use the key words from a small set of known relevant papers. This led to a very high false positive rate. The final strategy is given in Table 50. This initial search yielded 26 papers with abstracts which appeared potentially relevant. These papers were ordered and evaluated. As a number of papers which were expected to be included were not, ISI Web of Knowledge (www.isiknowledge.com) was used to identify all papers either cited by, or citing the initial 26 identified papers. These supplementary searches yielded 14 and 27 extra papers of potential interest. Additionally, one other important paper was identified subsequently, and was included (Lancsar, et al., 2011). The limit of the search to Medline and Embase is a limitation. However, augmenting these search results with the studies that have either cited, or been cited by, the identified studies should be able to identify all relevant studies.

Existing attempts to estimate a SWF using stated preference data

The scale of the evidence on constructing a SWF using stated preference data in healthcare decision making is limited. Three notable exceptions are studies by Dolan and Tsuchiya (2009), Bleichrodt *et al.* (2005) and a recent study by Lancsar *et al.* (2011). There are a number of papers which attempt to quantify some aspect of the trade-offs between people that an average respondent is willing to make. These studies provide some evidence which might help to inform the investigation described in this chapter; however, they are somewhat problematic in that their investigation of single issues may mean they cannot identify the value of health gains to people with a certain characteristic independent of all others. Dolan *et al.* (2005) produce an excellent review of these studies, and conclude that

“The results from a systematic review of the literature suggest that QALY maximisation is descriptively flawed. Rather than being linear in quality and length of life, it would seem that social value diminishes in marginal increments of both. And rather than being neutral to the characteristics of people other than their propensity to generate QALYs, the social value of a health improvement seems to be higher if the person has worse lifetime health prospects and higher if that person has dependents. In addition, there is a desire to reduce inequalities in health. However, there are some uncertainties surrounding the results, particularly in relation to what might be affecting the responses, and there is the need for more studies of the general public that attempt to highlight the relative importance of various key factors.” (p.197)

These results are important in shaping the direction taken in this chapter. The non-linearity in quantity and quality of life is something which will be investigated, as will the characteristics of the individual receiving the health gain. The drivers of the results which Dolan *et al.* (2005) identify as uncertain are difficult to identify. If people (for example) favour health gains which accrue to men rather than women, is this because they intrinsically value the health of males more highly, or that they are making some assumption about other characteristics of males (such as shorter life expectancy) which are the actual drivers of the preference? Using the DCE methodology introduced in Chapters 3 and 4, and tested in Chapter 5, is clearly of potential value. If the DCE is appropriately designed, it identifies the impact of characteristics independently of all others, a characteristic of considerable value in a situation in which important characteristics are unlikely to be independent of one another.

Before showing how this might be done, I will focus on the three papers which attempt to estimate the SWF, rather than just one specific trade-off. As noted previously, the three leading examples of this are studies by Dolan and Tsuchiya (2009), Bleichrodt *et al.* (2005) and Lancsar *et al.* (2011).

In their study, Dolan and Tsuchiya consider a SWF with constant elasticity of substitution (CES):

$$W = \left[\alpha U_a^{-r} + \beta U_b^{-r} \right]^{\frac{1}{r}} : U_a, U_b \geq 0, r \geq -1, r \neq 0, \quad \text{Equation 78}$$

where W is social welfare derived from health, and U_a and U_b are the levels of health of two groups a and b . The parameter r represents the degree of aversion to inequality (or the convexity of the SWF). If $r = -1$, society is indifferent to inequality. As r increases, society becomes increasingly averse to inequality and the SWF becomes convex. The parameters α and β show the relative importance of the two groups in contributing to societal welfare. A set of symmetrical SWFs that conform to these characteristics are given in Figure 31 (assuming $\alpha = \beta$), and also in Figure 32 which relaxes the assumption that $\alpha = \beta$.

Figure 31: A set of symmetrical SWFs with constant elasticity of substitution assuming anonymity (i.e. $\alpha = \beta$)

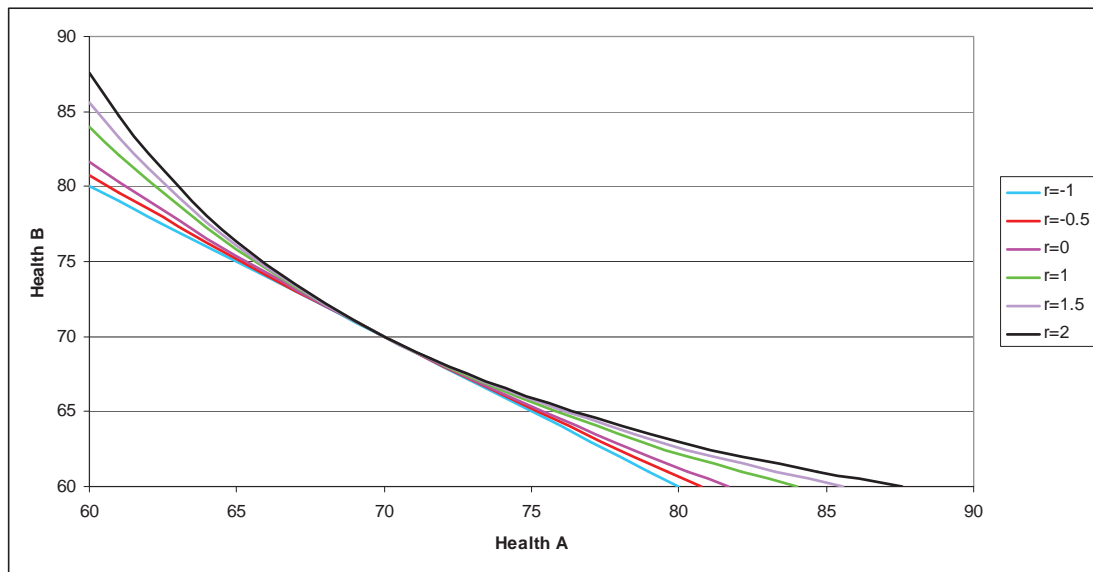
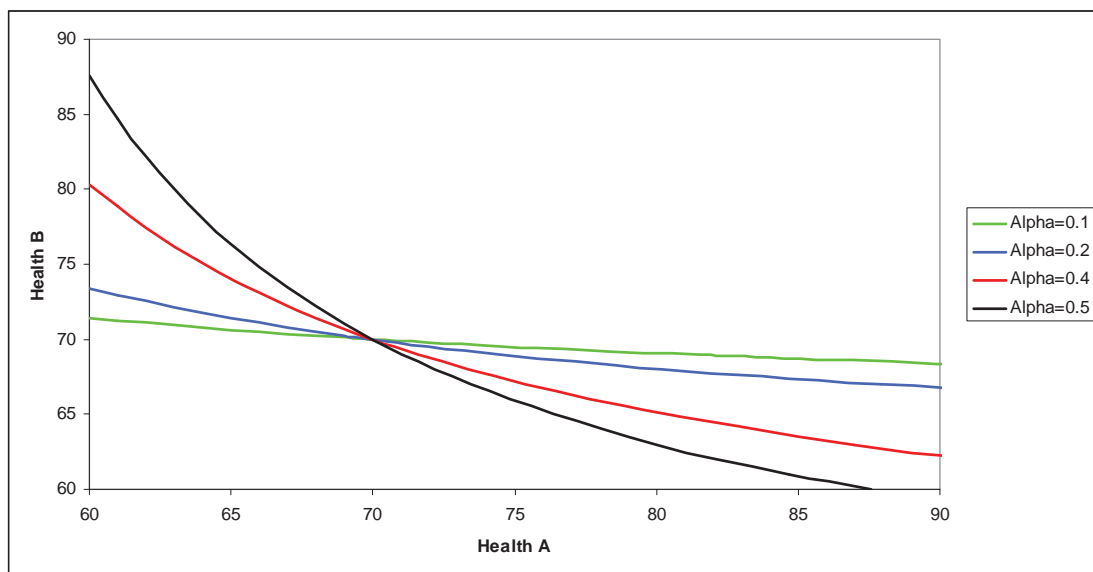


Figure 32: A set of SWFs with constant elasticity of substitution allowing differing interpersonal weights



As Dolan and Tsuchiya argue, estimating the value for r can be done if two points can be identified on a common SWF, and the value of α is known. If these two points are defined as $[X_A, X_B]$ and $[Y_A, Y_B]$, the marginal rate of substitution (MRS) at the midpoint of the two is

$$-\frac{dU_B}{dU_A} \Big|_{1/2(X+Y)} = \frac{\alpha}{(1-\alpha)} \left[\frac{U_B}{U_A} \right]^{(1+r)} = \frac{\alpha}{(1-\alpha)} \left[\frac{((U_B(X_B) + U_B(Y_B))/2)}{((U_A(X_A) + U_A(Y_A))/2)} \right]^{(1+r)} \quad \text{Equation 79}$$

We also know that (assuming that the two points are close)

$$\frac{dU_B}{dU_A} \approx \frac{U_B(Y_B) - U_B(X_B)}{U_A(Y_A) - U_A(X_A)} \quad \text{Equation 80}$$

By taking logs of these and solving for r , Dolan and Tsuchiya showed that

$$r \approx \frac{\log[(U_B(X_B) - U_B(Y_B))/(U_A(Y_A) - U_A(X_A))] - \log[\alpha/(1-\alpha)]}{\log[(U_B(X_B) + U_B(Y_B))/(U_A(X_A) + U_A(Y_A))]} - 1 \quad \text{Equation 81}$$

At the point on the SWF where $U_A=U_B$, the MRS is independent of the value of r (as r only impacts on the aggregate social welfare if there is an inequality between U_A and U_B). The approach Dolan and Tsuchiya took to estimating the value of α (the parameter which defines the SWF as symmetrical or otherwise) is to consider the impact of small changes away from the line of equality. Under one state, one group (called A) receives a health gain of p . Under another state, the other group (called B) receives a health gain of q . In both states, the health of the other group remains constant. If $p=q$ at the point of indifference between the two states, it can be inferred that $\alpha=\beta$. However, if $p>q$, it can be concluded that the health of group B is relatively more important in defining social welfare (and visa versa). The gradient of the tangent at the initial point of equality is approximately $-q/p$, and the gradient of the indifference curve will be

$$-\left. \frac{dU_A}{dU_B} \right|_{U_A=U_B} = -\frac{\alpha}{1-\alpha} \quad \text{Equation 82}$$

Combining the gradients allows an estimate of α

$$\alpha \approx \frac{q}{p+q} \quad \text{Equation 83}$$

The equity weight, defined as the marginal rate of substitution between A and B is then

$$MRS_{AB} \equiv \frac{dU_B}{dU_A} = \frac{\partial W}{\partial U_A} \frac{\partial U_B}{\partial W} = \frac{\alpha}{(1-\alpha)} \left[\frac{U_B}{U_A} \right]^{(1+r)} \quad \text{Equation 84}$$

This approach has considerable strengths: The estimation of relative importance of individuals can be estimated, as can the degree of inequality. However, there are two substantial limitations to their approach. Firstly, it is assumed that the median respondent is representative of the whole. This fails to account for strength of preferences. Secondly, it necessitates an assumption regarding functional form. While constant elasticity of substitution generates a range of plausible Social Welfare Functions, there is no convincing reason why this assumption should hold. There are a variety of reasons why a kinked SWF

might be plausible, for example at points where an individual reaches the often-assumed “Fair Innings” of seventy years (Williams, 1997).

Bleichrodt, Doctor and Stolk (2005) generate a SWF using the rank-dependent quality-adjusted life-year model. They consider a population of size n , and value the QALY profile of that population by

$$\sum_{i=1}^n \pi_i U(q_i) \quad \text{Equation 85}$$

where π_i is a utility weight placed on individual i , and $U(q_i)$ is the value society places on the quality-adjusted life-years received by individual i . The utility weight depends solely on the individual’s rank in terms of expected QALYs in the population. For an inequality averse decision maker, higher utility weights are placed on those whose rank is lower. The use of $U(q_i)$ rather than q_i allows for the utility function U over QALYs to be non-linear.

The approach Bleichrodt *et al.* (2005) adopt is two-fold. Firstly, they elicit the utility function with respect to QALYs. The purpose of this is to identify the diminishing marginal utility of QALYs independent of distributional impacts. Putting respondents in the position of decision maker, they asked respondents a series of questions to identify a point of indifference between two societal distributions of QALYs. Formally, they defined four QALY values x_1, x_0, R and r such that $x_1 > x_0 > R > r$. They then asked the respondent to choose between a proportion of the total population p receiving x_1 with the remainder $(1-p)$ receiving r , and an equal proportion p receiving x_0 with the remainder receiving R . If $x_1 - x_0 = R - r$, we would assume that the respondent would select the option in which p receives x_0 , and $(1-p)$ receives R as the curvature of the utility function of QALYs would assume diminishing marginal returns. As x_1 increases, it becomes increasingly likely that the respondent’s preference would switch. At the point of indifference, the task restarts with x_1 in place of x_0 and x_2 in place of x_1 . Using the relative values of all x_i terms, the slope of the utility function with respect to QALYs can be found.

The recent analysis by Lancsar *et al.* (2011) most closely matches the approach specified in Chapter 3, in that it uses a DCE, allowing simultaneous consideration of multiple areas of potential importance. The areas investigated by this study were the age of onset of disease, the expected age at death, the severity of the condition and the QALY gain associated with treatment. It is noteworthy that all of the dimensions of this DCE were specific to the disease or the treatment. Thus, while it does allow for non-symmetric preferences, it is somewhat

limited in that it does not consider non-health characteristics of the individual potentially receiving treatment. Whether this latter group of considerations should be included in any efficiency-equity trade-off is a judgement; however, the findings of Dolan *et al.* (2005) cited previously suggests that people do feel comfortable discriminating on those attributes.

Lancsar *et al.* (2011) conclude that one should not generally weight QALYs, suggesting that the conventional QALY model is fairly reflective of preferences. This is a correct interpretation of their data, but might reflect some specific characteristics of their experiment, most notably the choice of dimensions and levels in their experiment, and their choice of method for deriving welfare measures. The first of these issues will be left until the discussion of this chapter as it is more interesting to compare the selection of attributes and levels at a point where results from my DCE can be contrasted with those of Lancsar *et al.* (2011). The issue of deriving welfare measures from DCE data was discussed in Chapter 3, and will be briefly outlined below.

To identify the willingness of survey respondents to trade off health between different groups, Lancsar *et al.* (2011) adopt the Hicksian compensating variation (CV). This is defined as

$$CV = \frac{1}{\lambda} \left[\ln \sum_{j=1}^J e^{V_j^0} - \ln \sum_{j=1}^J e^{V_j^1} \right], \quad \text{Equation 86}$$

where V_j^0 and V_j^1 are the value of the representative indirect utility function for each choice option j before and after the change of interest; J is the number of options in the choice set; and λ represents the marginal utility of the chosen numeraire, often income but in this case the marginal utility of a QALY. The change of interest in this situation is the move from the base case respondent to a different respondent. The CV was described diagrammatically in Chapter 3 (Figure 12).

The equity weight (termed distributional weight by Lancsar *et al.*) is then estimated as

$$\text{Weight} = 1 - \frac{CV}{QALY_{base}}, \quad \text{Equation 87}$$

where $QALY_{base}$ is the number of QALYs gained in a pre-specified reference case and CV is the number of QALYs required to equalise expected utility between profiles V_j^0 and V_j^1 . A

positive CV leads to a weight less than 1 (i.e. that health gain for this group is valued less than health gain for the reference group).

Identifying dimensions for the DCE

Olsen *et al.* (2003) identify a range of characteristics of an individual which may affect how society values health gains accruing to them. A systematic review of the literature identified characteristics in three broad areas which may be of relevance. These areas, and the characteristics identified as potentially relevant, are outlined in Table 39. The listed characteristics are certainly not mutually exclusive, nor are they exhaustive. They summarize the terms that have been applied in the identified studies, and represent a pool from which possible individual-level characteristics of interest might be derived.

Table 39: Potentially relevant personal characteristics identified by Olsen *et al.*

A person's relation to others	A person's relation to (the cause of) the illness	A person's self
Marital status	Contributed to illness	Gender
Have children	Have taken care of their own health	Sexuality
Caring for elderly relative	Self-inflicted	Race
Breadwinner	Smoker (vs non-smoker)	
Unemployed	Unhealthy diet	
Unskilled (vs director)	Poor diet vs inherited disease	
Lorry driver (vs teacher)	High alcohol consumption	
Important to the community	Illegal drug use	
Employed	Rarely exercise	
Rich		
Lower socio-economic status		
Deprived		
Criminal record		

Olsen *et al.* (2011) deliberately excluded age (or life expectancy or quality-adjusted life expectancy) from their review as they argued “*we consider age to be related to a point in each person's life-time as distinct from characteristics which separate one individual from another*”. This distinction is less obvious than Olsen *et al.* (2011) make it seem, as a number of the characteristics they have included are also strongly associated with age. This group would include being a breadwinner, having dependent children or being important to the community.

Mooney *et al.* (1995) investigated the relative importance of groups of people in a welfare function (but that of health care professionals and decision makers rather than the general population). In a sample of 283, they identified that this population had preferences in favour of allocating health gain to the young, those with poor health or low socioeconomic status (SES), for health gains that occur more immediately, and for splitting a fixed health gain between a larger set of recipients. This analysis is valuable insofar as it considers a range of potentially important factors influencing a social welfare function. However, it also highlights a significant flaw which can affect studies attempting to consider preferences which are asymmetric around the 45 degree line in the social welfare function. It is unclear whether the preferences for particular groups are preferences for that group *per se*, or whether the respondent is making assumptions regarding the other characteristics of the group. For example, is the respondent assuming that people with poor health are of a lower SES? If so, identifying that the group prefers allocating health gain to people with poor health may not be because the respondents are interested in poor health, but rather that they are interested in SES. While it is implausible that the respondents had preferences as rigid as this (in terms of having no interest in poor health independent of SES), it highlights that a study aiming to quantify the relative importance of various individual-level characteristics needs to be able to identify the effect of one dimension independent of all others.

Including gender

Tsuchiya and Williams (2005) discuss the “fair innings” argument between genders, both whether an inequality exists in favour of women and whether such an inequality can be considered as inequitable. They put forward a series of reasons why gender inequality may not be a convincing candidate for being labelled inequitable, and conclude these to be flawed. Of the reasons they dismiss, they are least certain about dismissing the notion that relatively short life expectancy in males reflects the choice by the average male to make more risky decisions. They note the argument of Le Grand (1991), who argues that an inequality caused by the deliberate choice of an informed individual will be much less inequitable than other inequalities. However, this argument is founded on the concepts of what constitutes deliberate choices and informed individuals. This clear distinction between individual responsibility and fortune was outlined by Dworkin (1981a; 1981b), who makes the distinction between an individual’s preferences and his resources. The difficulty in applying this concept in practice lies in the significant grey area between the two. One counter-argument to asserting that the gender inequality does not represent an inequity

suggested by Tsuchiya and Williams is to argue that the willingness to accept risk of men is an evolutionary necessity and therefore it is not reasonable to judge the inequality as equitable. In my view, neither extreme view is satisfactory. While risk-taking activity is associated with the interaction between society and gender, individual responsibility cannot be consumed by biological imperatives.

The conclusion that Tsuchiya and Williams (2005) reach that positive discrimination in favour of men is warranted is based on the idea that men do have a lower life expectancy than women. While this appears to be true in most populations, the purpose of including gender in an experiment would be to ascertain if one gender is favoured independent of any difference in life expectancy (quality-adjusted or otherwise). While a priori such a conclusion would not be expected, the evidence on this has actually identified some weak trends in favour of both sexes (Charny, et al., 1989; Dolan, et al., 1999; Mooney, et al., 1995). As argued previously, the strength of analysing this issue in the context of a DCE is that it allows identification of the effect of gender independent of other factors which are explicitly stated. Thus, any assumption that, for example, women live longer so should receive a lower priority, would be captured by a population's aversion to inequality of life expectancy rather than in an aversion to women. The other potential reason for gender to be informative is that, while it is unlikely the average individual would allocate healthcare resources on the basis of gender, sub-group analysis may identify heterogeneity in this dimension.

Age weighting

The issue of considering age as a factor in weighting healthcare-derived outcomes has been addressed by a number of studies (Bognar, 2008; Johannesson and Johannesson, 1997; Nord, et al., 1996; Rodriguez and Pinto, 2000). As noted by Rodriguez and Pinto (2000), discrimination by age can be driven by either efficiency or equity. The efficiency-based argument claims that greater weight should be given to those of working age as they contribute more to social welfare (Murray and Lopez, 1994). In adults, conclusions derived from this principle largely agree with those derived from the type of equity based argument asserted by Williams (1997). The area in which they diverge is the treatment of children, as they are currently socially unproductive (although will go on to become so), whilst being the furthest from receiving a Fair Innings. What is important for this work is that there are arguments in favour of age discrimination, and evidence that society is willing to do so.

Therefore, including age in the choice experiment seems an important addition despite its exclusion from Olsen's (2003) list of relevant personal characteristics.

One important issue which is raised by Rodriguez and Pinto (2000) is the possibility that age weights between individuals are not constant as the size of the gain changes. For example, while society may be indifferent between 10 years for a 20-year old and 5 years for a 60-year old (implying a weight for the 60-year old of 0.5 relative to the younger person), it does not necessarily follow that society would be indifferent between 40 years for the 20-year old and 20 years for the 60-year old. Indeed, the Fair Innings gives us a good explanation why the age weight might differ depending on the scale of the gain. If reaching the Fair Innings is of intrinsic value, years added beyond the Fair Innings may be of lower societal value than those accrued in order for an individual to receive it. Rodriguez and Pinto (2000) provide evidence regarding this variability of age weight. For small gains in life expectancy, the weight placed on 40 year-olds is comparable to that placed on 20 year-olds. However, when considering a gain of 40 years, it is valued considerably lower in the 40 year-olds (73% of the value given to the same gain in 20 year-olds). The pattern is more pronounced when considering the weight placed on 60 year-olds. The first year is valued at 70% of the value of an additional year for a 20 year-old, but this decreases to 63% for a health gain of 10 years, and 55% for 20 years. One caveat which should be added to this study was that the population they used was small ($n=61$) and consisted of undergraduate students (many of whom would be close to 20 years old). While Rodriguez and Pinto (2000) do not claim that it is representative of the entire population, others have ignored this issue when considering the impact of introducing age weights (Eisenberg and Freed, 2007).

Life expectancy, current age or both?

One consideration for the design of the choice experiment is how to frame the group's current endowment of health, and their expectation of future health with and without the intervention. Possible dimensions in the experiment include life expectancy, quality-adjusted life expectancy, current age or some combination of two or more of these. This is important for both data analysis (and the type of conclusions that can be drawn from the results) and also for the size of the experiment (and hence the sample size) required to estimate the effects that might a priori be considered interesting and important.

If equity is a concern, expected lifetime is clearly important for age-weighting. Williams (1997) goes one step further by arguing that expected lifetime should be adjusted to account

for morbidity, therefore using quality-adjusted life expectancy. However, the weights generated by Murray and Lopez (1994), and also by Rodriguez and Pinto (2000) are based on current age rather than (quality-adjusted) life expectancy. It is arguable that both should be considered in an experiment. If a rationing decision has to be made between a twenty-year old and a sixty-year old, both expected to live to seventy, it is not clear on equity grounds which should be preferred. However, including both age and life expectancy poses significant issues in terms of generating impossible combinations of the two as life expectancy must exceed current age. The approach Williams uses is to present the survey respondent with two groups in society with particular quality-adjusted life expectancies (and no current age).

Selecting dimensions and levels for the DCE

In this chapter, the personal characteristics that have been suggested and investigated as determinants of societal preferences for healthcare decision making have been outlined. Evidence of willingness to discriminate in these dimensions is present in most cases, although the methods quantifying these trade-offs have been questioned. In particular, with one exception, studies have not been built to consider the impact of a characteristic independent of all others (Schwappach, 2003). In designing a choice experiment to investigate the issue, a number of concerns have to be balanced. While, it is important that the most significant of these personal characteristics are captured and assessed, it is also necessary to limit the number of characteristics (and levels of each characteristic) so the experiment is of a manageable size, and given a finite sample size, that the choice sets are adequately populated to quantify the impact of each characteristic with confidence.

The major source of dimensions with supporting evidence was the review by Olsen *et al.* (2003) which suggested a number of possible dimensions that might be important dividing into those relating to a person's relation to others, those relating to their illness, and those relating to their self. From these, a smaller set were selected with the aim of including some from each of these three categories. The selected dimensions were gender, smoking status, income (or socio-economic status), whether the individual maintained a healthy lifestyle, carer status and total life expectancy. These are not an exhaustive set of characteristics over which people might discriminate, only that these are a convenient and obvious set which can help to identify the degree to which people agree or disagree with the standard health-maximising approach. Therefore, the results presented here should not be interpreted as

claiming that there are no other characteristics which might impact on preferences, only that, over a set of obvious candidates, respondents either do or do not correspond to the assumptions of the QALY model.

From the findings of the literature review, and the issues relating to minimising respondent burden, the following dimensions given in Table 40 were selected for the choice experiment.

Table 40: Dimensions and levels for the choice experiment

Dimension	Levels	Detail and coding
Gender	2	Male=0, Female=1
Income	2	Below average=0, Above average=1
Smoking	2	No=0, Yes=1
Healthy lifestyle	2	No=0, Yes=1
Carer status (dependents)	2	No=0, Yes=1
Group life expectancy (years)	4	30=0,45=1,60=2,75=3
Gain in life expectancy (years)	4	1=0,3=1,6=2,10=3

Four levels were selected for both the initial life expectancy and the gain in life expectancy. The choice of four was based on the relative ease of design and efficiency of experiments with dimensions all having a common root (i.e. that all dimensions are powers of the same prime, in this case 2). The choice of initial life expectancy and gain in life expectancy involved a balance between plausibility of health gain for particular groups, and capturing a broad spectrum of ages and outcomes. Mæstad and Frithjof Norheim (2009) presented an argument why small gains in life expectancy should be used to estimate the age weight at a point. When comparing the relative importance of life at two ages x and y , stated preference experiments offering life extensions beyond those ages rely on the valuation of health in the years $x+\delta x$ and $y+\delta y$. Therefore, ten years was selected as the maximum potential gain despite there being a range of interventions that might conceivably increase life expectancy, particularly among the young, by a much longer period.

An example choice set is given in Figure 33.

Figure 33: An Example Choice Set

CHERE
CENTRE FOR HEALTH ECONOMICS
RESEARCH AND EVALUATION

If you were asked to choose one of the following two programs, each of which would impact on the health of 100 people, which would you select?

	Program One	Program Two
The people in this group are	Male	Female
The people in this group have an income which is	Above average	Above average
The people in this group are	Smokers	Non-smokers
In terms of diet and exercise the people in this group have a lifestyle which is generally	Healthy	Healthy
Are the people in this group fulltime carers (e.g. for children or for adults with medical conditions)?	Yes	No
Without the program, the people in the group will live until they are	30 years	30 years
The program would increase their life expectancy by	1 year	10 years
Which program would you choose?	<input type="radio"/>	<input type="radio"/>

progress

Designing the choice experiment

For the experiment, a design containing 5 2-level attributes and 2 4-level attributes was required. A starting design of 2^5 of strength four in 16 rows (i.e. half of the full factorial is available). This is reproduced in Table 41.

Table 41: A starting design of 2^5 in 16 rows (strength 4)

0,0,0,0,0	0,0,0,1,1	0,0,1,0,1	0,0,1,1,0
0,1,0,0,1	0,1,0,1,0	0,1,1,0,0	0,1,1,1,1
1,0,0,0,1	1,0,0,1,0	1,0,1,0,0	1,0,1,1,1
1,1,0,0,0	1,1,0,1,1	1,1,1,0,1	1,1,1,1,0

If a design is strength four, it means that, within any four dimensions, each possible combination of levels occurs with equal frequency. A design of strength four allows for estimation of both main effects and all two-factor interactions (Street and Burgess, 2007). Each of these rows was then paired with each possible combination of the two four-level attributes (group life expectancy and treatment gain), giving a design with 256 rows.

The next stage was to identify a set of generators which allowed estimation of all main effects and interactions. For a generator to estimate the main effect of a dimension being a

particular level, the primary condition must be that it must be non-zero. In addition, a generator for an attribute should not contain a non-trivial divisor of the number of levels within that attribute (e.g. an eight-level attribute should not have two or four as the generator). The reason for this is because it would not provide a complete ranking. For example, if an eight-level attribute is coded 0-7, and a generator of 2 is applied to each of the levels, it would be possible to discern the relative value placed on levels 0,2,4 and 6, and to the relative value placed on levels 1,3,5 and 7, but no comparison between these sets of levels would be completed. For our four-level attributes, this means that the generators should consist solely of 0s and 1s (it might also be useful to allow 3s in the generators, but this was not done in this design).

To identify a suitable set of generators, the approach advocated by Street and Burgess (2007) was taken. As stated in Chapter 3, a suitable set of generators is such that

- For each attribute, there is at least one generator with a 1 in the corresponding position, and
- For any two attributes there is at least one generator in which the corresponding positions have a 0 and a 1. (p.129)

Using Street and Burgess's example 4.2.8 (for which there are eight dimensions rather than the seven required here), the following generators were considered:

1,1,1,1,0,0,0,0

1,1,0,0,1,1,0,0

1,0,1,0,1,0,1,0

0,1,0,1,0,1,0,1

The property that makes these generators appropriate is that a 1 occurs at least once in each column and, for each pair of columns, there is a row with both a 1 and a 0 in at least one of the generators. As these have eight columns and there are seven attributes in the experiment, the final column can be removed. Within the seven columns in the first three generators, all main effects and two-factor interactions are estimated: therefore, the final generator can also be removed, leaving the following generators:

1,1,1,1,0,0,0

1,1,0,0,1,1,0

1,0,1,0,1,0,1

When applied to the 256 rows of the orthogonal array, these generators produce a design with 640 choice sets: these are provided in Appendix 9. Note that the 640 choice sets is a reduction on the $256 \times 3 = 768$ that might be expected given the number of rows and generators. This is because there is duplication of pairs, which can be removed.

The Λ -, B -, and C - matrices are very large, and hence not reproduced here. The Street-Burgess software (<http://maths.science.uts.edu.au/math/wiki/SPExptSoftware>) reported that no effects were correlated.

Sample recruitment

This was subject to a pilot (N=241, reported in Norman and Gallego (2008)), which concluded that respondents generally found the task straightforward and comprehensible. The main data collection occurred in May 2010. An online panel of respondents was used for the survey recruited by Pure Profile Pty. These respondents were each paid a small sum (approximately \$15) to complete the survey. To allow comparability with the Australian population, respondents were selected according to age and gender. Each respondent used a web link to access the survey, so were able to self-complete at their convenience. To aid the respondent, a thorough description of the task was provided at the beginning of the survey and a help button was available throughout the task. This provided information on how to respond. They then completed the task for the 16 choice sets. Following this, they answered a series of personal questions including gross household income, ethnicity, country of birth, number of dependents, level of education, age and gender. Finally, they were asked how difficult the task was, selecting one of five levels of difficulty ranging from very difficult to very easy. They were also given the opportunity to provide a free-text response outlining their impression of the survey.

Screenshots of each of the pages within the experiment are provided in Appendix 10.

Analysis

With some exceptions, the analysis followed the strands described in Chapter 3. Thus, fourteen models were attempted. Models A and A1-A6 imposed a QALY-type model in which the utility of alternative j in scenario s for individual i is

$$U_{isj} = \alpha GAIN + \beta X'_{isj} GAIN + v_i + \varepsilon_{isj} \quad \text{Equation 88}$$

This, the sole main effect was on the health gain, and the other characteristics enter the utility function as interactions with the health gain attribute. As noted in Chapter 3, this was an important amendment to a model using main effects on all levels of interest as it imposes the zero-condition in which all options where health gain was zero were valued equally irrespective of the characteristics of the hypothetical person ‘receiving’ it.

The models A1-A6 which were paired with this utility function, with their underlying strengths and weaknesses are described in detail in Chapter 3, and were:

- A1: The Conditional Logit
- A2: The Scale Multinomial Logit
- A3: The Mixed Logit (Uncorrelated Coefficients)
- A4: The Generalised Multinomial Logit (Uncorrelated Coefficients)
- A5: The Mixed Logit (Correlated Coefficients)
- A6: The Generalised Multinomial Logit (Correlated Coefficients)

Relaxation of the utility function (models B, B1-B6)

In models A1-A6, I have considered a utility function which is linear with respect to gain in life expectancy (and when coupled with the zero condition, this forms the QALY model). This is a strong assumption, and requires testing. Utility Function B builds on Utility Function A by relaxing the assumption of linearity of utility with respect to time. Thus, Utility Function B is:

$$U_{isj} = \alpha GAIN + \delta GAIN^2 + \beta X'_{isj} GAIN + \phi X'_{isj} GAIN^2 + v_i + \varepsilon_{isj} \text{ Equation 89}$$

Thus, the linearity of utility with respect to time is relaxed, as reflected in the $\delta GAIN^2$ term in Equation (89). In addition, it relaxes the assumption that the change in total utility associated with it being received by a different group of hypothetical respondents is independent of the total gain (the $\phi X'_{isj} GAIN^2$ term)

Models B1-B6 are therefore replications of models A1-A6 respectively, but adopting this more relaxed utility function. Note that model B requires a large number of additional parameters, and that the number of additional parameters increases substantially as I move away from the more restrictive models. As described in Chapter 3, model evaluation will primarily be undertaken using the Akaike and Bayesian information criteria (AIC and BIC)

(Akaike, 1974; Schwarz, 1978). These consider both the model fit and also the parsimony of the model (by accounting for the number of parameters in the model).

Models B and B1-B6 are therefore replications of models A and A1-A6 respectively, but adopting this more relaxed non-linear utility function. A summary of the various models run is provided in Table 42.

Table 42: Models Run in Chapter 6

		Utility Function 1	Utility Function 2
	RE Probit / Logit	A	B
Heterogeneity modelling	Conditional logit	A1	B1
	Scale MNL	A2	B2
	Mixed logit	A3	B3
	G-MNL	A4	B4
	Mixed logit (correlated)	A5	B5
	G-MNL (correlated)	A6	B6

Self-interest and empathy – sub-group analysis

One additional question which can be answered is the degree to which people value health outcomes that accrue to people with characteristics similar to themselves. The demographic data collection allows the respondent to be described under each of the binary variables in the experiment (smoking, healthy lifestyle, carer status, income and gender). While it is not possible to identify an individual's life expectancy, this will be somewhat related to current age. Thus, once a preferred model was identified, the analysis was replicated with sets of respondents defined by each of these categories. The null hypothesis was that the (for example) gender of the respondent does not impact on the valuation the respondent places on health gains that accrue between genders. However, it is plausible that this null is incorrect; therefore a full sub-group analysis was undertaken. This was done using the base case RE probit approach. This type of analysis has previously been undertaken in the context of the mixed logit (with the impact of observables on mean coefficients assessed) (Harris and Keane, 1999). The reason for using the base case approach rather than the more flexible mixed logit or G-MNL was that the extra parameters needed to undertake this sub-group study meant that convergence in all models would be difficult to achieve, and take substantial time. In addition, interpreting differences in coefficients beyond those in the base case becomes increasingly difficult as constraints are relaxed.

The approach taken to sub-group analysis was to re-estimate the base case, but with an additional coefficient for each dimension reflecting the interaction between gain in life expectancy, each characteristic of the hypothetical individual, and a particular demographic characteristic. Thus, the utility function of alternative j for individual i with or without demographic characteristic c in scenario s is

$$U_{icsj} = X'_{isj}(\beta_i + \beta_c) + \varepsilon_{isj} \quad \text{Equation 90}$$

Each possible c (smoking status, healthy lifestyle, gender, above average income, carer status, and above average age) were run separately, giving five regressions. Gender, healthy lifestyle, above average income and carer status were binary (yes or no); therefore, these regressions had one extra term for each coefficient in the regression relating to the ‘Yes’ option for each. Smoking had three options in the demographic data collection stage (never regularly smoked, former smoker and current smoker). Therefore, two extra terms were generated for each coefficient with never smoked the omitted level. To explore whether allowing for different responses based on each characteristic of the respondent, a likelihood ratio test was used for each in turn. In principle, it would be possible to explore response patterns across multiple dimensions of the respondent simultaneously. For example, it would be possible to investigate if male smokers differ from other groups (i.e. male non-smokers, female smokers, female non-smokers). However, the sample size collected in the experiment means that the number of observations driving each coefficient in this larger regression would be increasingly small making reliable inferences difficult.

The demographic questions asked in the survey allowed the option for the respondent to decline to disclose. This has some implication for this analysis if types of respondents were more likely to decline response. The number of decliners is noted in the respective results.

Generating equity weights from regression results

The previous chapter discussed the appropriate techniques for generating utility weights for generic quality of life instruments, and a similar technique can be applied in this context, albeit with some notable differences as a result of different approaches required to anchor values and a different omitted level in the regression.

In the context of utility weights for economic evaluation, I wanted to anchor weights such that health states considered to be as bad as death were assigned a value of zero, while states equivalent to full health were assigned a value of one. Importantly, the experiment had to be

designed in such a way that health states valued in the range between zero and one had the important trade-off principle that was discussed in the introductory chapter of this thesis, and elsewhere (Flynn, 2010). Notably, an average individual experiencing a health state with a value of (for example) 0.5 would be willing to trade-off half of their remaining life expectancy to be returned to full health. To make this possible, Flynn argued that life expectancy was a necessary inclusion in the choice experiment.

In the context of equity weights, anchoring is somewhat different. It seems sensible that a health gain accruing to an average individual should have an equity weight of one. This would mean that a health intervention applied to the entire population would produce the same result whether we apply equity weights or not. As I have previously stated that equity in this chapter is focused on equality of outcome, this is correct. Equity weights would then be unbounded with an equity weight of greater than one applied to a group who the choice experiment data suggest should receive additional emphasis, a weight of between zero and one for gains in groups which the data suggests are valuable but less so than average, and a weight of less than zero for gains in groups that the data suggests should be valued negatively.

Equation (88) gave the more restrictive utility function specification described previously. This is replicated below:

$$U_{isj} = \alpha GAIN + \beta X'_{isj} GAIN + \varepsilon_{isj} \quad \text{Equation 91}$$

To reiterate this, the systematic utility to individual i of option j in choice set s of a health gain to a particular hypothetical group is modelled as consisting of a main effect on the gain to the group plus interactions between the gain and the characteristics of the hypothetical group receiving the health gain (gender, smoking status etc). Dropping the subscript for simplicity, this can be differentiated with respect to $GAIN$:

$$\frac{dU}{dGAIN} = \alpha + \beta X' \quad \text{Equation 92}$$

Up to this point, the analysis is identical to that seen in Chapter 5. However, the anchoring required in this context is different. In this chapter, the equity weight for a hypothetical group with particular characteristics, X' , is estimated by dividing through by the value for the population mean, i.e.

$$\frac{\alpha + \beta X'}{\alpha + \beta \bar{X}}$$

Equation 93

For this analysis, it is of course necessary to define what is meant by an average member of society. For the purposes of demonstrating the method, a simple mean member of society was assumed. As society divides approximately in half in terms of gender, the mean respondent was assumed to fall halfway between the two. Therefore, the β term applicable was a midpoint between the male coefficient (0 as it was omitted from the regression) and the female coefficient. For convenience, this reference group was selected to be the ‘average’ group in society, under the assumptions that 50% of people in society are female, that 50% have above average income, that 50% have a healthy lifestyle, that 20% are smokers, that 40.8% are carers, and that the average person has a total life expectancy of 75). The carer figure is a composite term including the 2.6 million Australians estimated by the Australian Bureau of Statistics to provide assistance to those who needed help because of disability or old age, the 2.363 million couple families with children (so 4.726 million parents) and the 1.944 million single parents (both parenting statistics are taken from the 2006 census (Australian Bureau of Statistics, 2006b)), divided by the estimated total population as of 12th October 2011 of 22.731 million (Australian Bureau of Statistics, 2011). Note that this mean respondent was simply the respondent with the mean characteristic in each dimension (i.e. that characteristics are independent of one another). This is a simplification and may lead to a mean equity weight other than one; however, in the absence of appropriate data to allow the independence assumption, this is a necessary step.

Therefore, for the mean respondent, $X' = \bar{X}$, and Equation (93) equals 1. This approach uses the respondent utility of extra health for the mean hypothetical individual or group as the numeraire. As Equation (93) divides one function of coefficients by another, the issue of scale is accounted for and the result shows the trade-off between health gains accruing to different hypothetical individuals. While slightly less elegant than the solution used in the utility weights chapter, it is fundamentally the same approach. To explore the degree of certainty in the results, confidence intervals for each of the equity weights were bootstrapped using 50 replications.

As in Chapter 5, this method can be easily adapted to generate weights under the more relaxed utility function allowing utility to be non-linear in time. Thus, the more relaxed utility function is

$$U_{isj} = \alpha GAIN + \delta GAIN^2 + \beta X'_{isj} GAIN + \phi X'_{isj} GAIN^2 + v_i + \varepsilon_{isj} \text{ Equation 94}$$

This can then be differentiated as in Equation (92), and equity weights generated as in Equation (93), with the corresponding ratio being

$$\frac{\alpha + \beta X' + 2GAIN (\rho + \phi X')}{\alpha + \beta X' + 2GAIN (\rho + \phi \bar{X})} \text{ Equation 95}$$

Results

Seven hundred and forty nine people entered the survey and were eligible to participate. Thirty-two of these were excluded as the sample had reached its maximum quota (i.e. a pre-specified quote was determined based on our budget, and 32 potential respondents clicked on the link after that number had been reached). Of the remaining 717, 616 answered at least one choice set (i.e. they did not withdraw before the task began) Of these, 553 completed all choice sets within the survey, giving a completion rate of 89.8% relative to those that started the task (and were therefore randomised to a block), and 77.1% relative to the population who entered the task and were willing to participate. Of these 553, one respondent completed the choice task (and formed part of the analysis set) but did not complete the demographic section at all. The characteristics of the sample of 552, and its comparability to the general Australian population are outlined in Table 21.

Table 43: Representativeness of DCE Sample

Characteristic	Value / Range	Sample	Population ¹
Gender	Female	56.16%	56.09%
Age (years)	16-29	26.63%	21.33%
	30-44	34.96%	23.98%
	45-59	23.01%	22.40%
	60-74	11.05%	14.00%
	75+	0.54%	18.29%
Highest level of education	Primary	3.26%	40.51%
	Secondary	30.43%	20.00%
	Trade certificate	30.43%	22.24%
	Bachelor's degree or above	35.87%	17.26%
Gross household income ¹	<\$20,000	7.84%	15.77%
	\$20,000 - \$40,000	15.88%	23.02%
	\$40,001 - \$60,000	20.59%	17.64%
	\$60,001 - \$80,000	17.84%	13.87%
	\$80,001 - \$100,000	15.29%	11.03%
	\$100,001 +	22.55%	18.67%

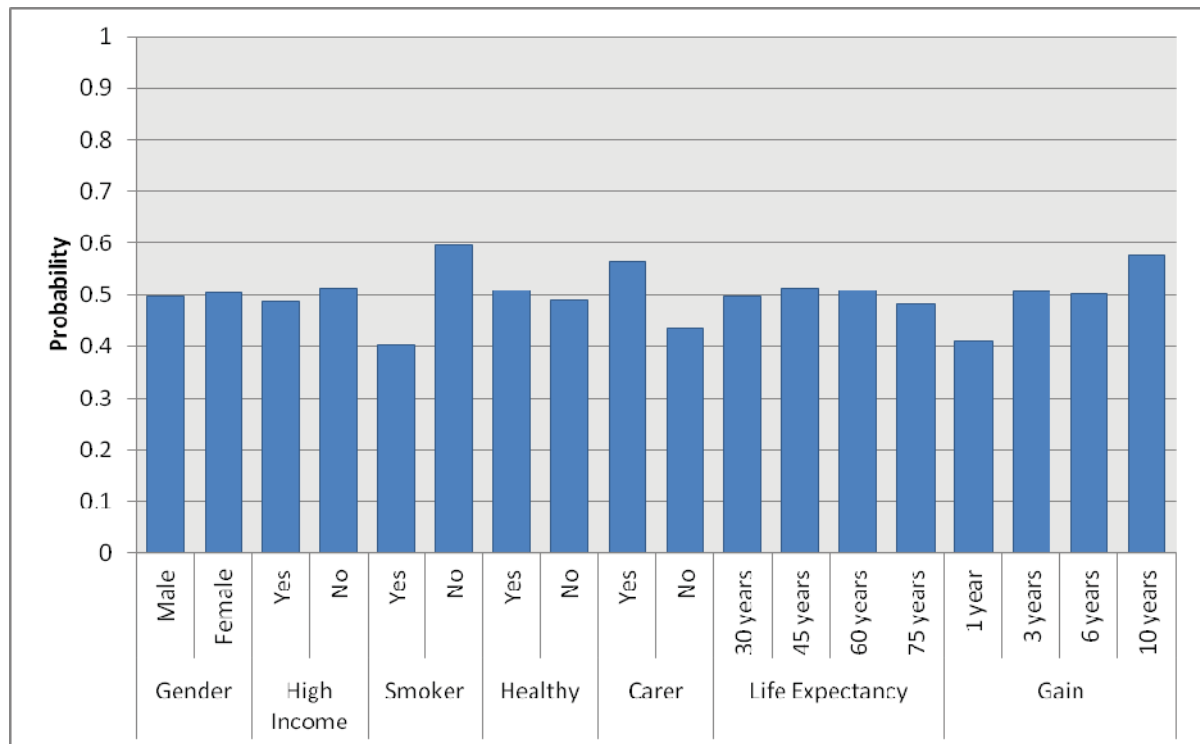
¹ All data sourced from ABS (Australian Bureau of Statistics, 2006a; Australian Bureau of Statistics, 2002; Australian Bureau of Statistics, 2005; Australian Bureau of Statistics, 2007)

The representativeness of the sample differs by characteristic. The gender breakdown is close to the population. Those over 75 years old are under-represented, which is a problem for generalisability. People in the sample are relatively over-educated and have a higher income than average.

Marginal frequencies

The marginal frequencies for each level are shown in Figure 34.

Figure 34: Marginal Frequencies



The marginal frequencies suggest that gain in life expectancy, smoking status and carer status are the factors that most influence the decision to choose an option. There is an average preference in favour of females, individuals with a low income, with a healthy lifestyle, or with a current life expectancy in the middle range of 45-60 years. *A priori*, it might be expected that the marginal frequency on a current life expectancy of 30 years might be higher; this will be explored in the context of heterogeneity.

Considering choice sets in which the gain differed between groups, the proportion in which the option producing the fewer years of additional life expectancy was selected was 32.3%. Thus, gain is important, but not the sole determinant of choice in the experiment. Similarly, of the 553 complete respondents, 106 never selected an option involving the fewer number of additional years of life. This means that the remaining 447 were willing to trade aggregate life years in order to focus health gain towards specific members of society.

One notable result is that the likelihood of selecting a program with a gain of 3 years appears to be higher than that of selecting a program with a gain of 6 years. However, this result is an artefact of the design approach. In the choice of generators, the values that applied to the Gain dimension were either zero or one. The absence of a two was justified as using a factor of the total number of levels in a dimension as a generator leads to non-complete rankings.

What this meant in the choice sets was that an option with ‘6 years’ in one of the hypothetical programs was only paired with a program with 3, 6 or 10 years (i.e. it is paired with 6 years if the generator is 0, and 3 or 10 years if the generator is 1). Similarly, for options with a gain of 3 years, the other option could take the value of 1 year, 3 years or 6 years. This explains why 3 years was actually selected more frequently. In situations in which 3 years and 6 years were the two gains visible for the two options in the choice set, the option with a 3 year gain was only selected in 35.7% of these choice sets. Running an additional conditional logit (not presented) with gain dummy coded and the only variable included confirmed this pattern.

Random-Effect probit results

The RE probit results, both under the assumption of linearity (Utility Function A) and under the more relaxed model (Utility Function B), are presented in Table 44.

Table 44: RE Probit Results

Mean (standard error)	Utility Function A	Utility Function B
Constant	-0.0350 (0.0139)**	-0.0351 (0.0140)**
Gain (years)	0.1092 (0.0068)***	0.2089 (0.0282)***
Gain x female	0.0035 (0.0024)	-0.0043 (0.0095)
Gain x high income	-0.0079 (0.0028)***	-0.0252 (0.0103)**
Gain x smoker	-0.0739 (0.0033)***	-0.1851 (0.0136)***
Gain x healthy life	0.0154 (0.0046)***	0.0487 (0.0163)***
Gain x carer	0.0317 (0.0027)***	0.1041 (0.0108)***
Gain x LE45	0.0140 (0.0054)**	0.0240 (0.0198)
Gain x LE60	0.0097 (0.0062)	0.0347 (0.0223)
Gain x LE75	-0.0094 (0.0055)*	-0.0211 (0.0199)
Gain ² (years)		-0.0096 (0.0027)***
Gain ² x female		0.0009 (0.0011)
Gain ² x high income		0.0020 (0.0011)*
Gain ² x smoker		0.0139 (0.0016)***
Gain ² x healthy life		-0.0037 (0.0017)**
Gain ² x carer		-0.0088 (0.0013)***
Gain ² x LE45		-0.0011 (0.0021)
Gain ² x LE60		-0.0027 (0.0023)
Gain ² x LE75		0.0013 (0.0021)
Lnsig ² u	-13.2502 (10.0324)	-13.5833 (11.083)
Sigma u	0.0013 (0.0067)	0.0011 (0.0062)
Log likelihood	-5570	-5496

Levels of statistical significance: *=10%; **=5%; ***=1%

The corresponding analysis using the *xtlogit* command is presented in Table 45.

Table 45: RE Logit Results

Mean (SE)	Utility Function A	Utility Function B
Constant	-0.0566 (0.0227)**	-0.0573 (0.0229)**
Gain (years)	0.1842 (0.0117)***	0.3409 (0.0467)***
Gain x female	0.0062 (0.0040)	-0.0062 (0.0155)
Gain x high income	-0.0130 (0.0046)***	-0.0414 (0.0168)**
Gain x smoker	-0.1233 (0.0057)***	-0.3012 (0.0224)***
Gain x healthy life	0.0264 (0.0079)***	0.0819 (0.0272)***
Gain x carer	0.0520 (0.0045)***	0.1692 (0.0176)***
Gain x LE45	0.0227 (0.0089)**	0.0411 (0.0323)
Gain x LE60	0.0154 (0.0102)	0.0589 (0.0365)
Gain x LE75	-0.0159 (0.0090)*	-0.0343 (0.0325)
Gain ² (years)		-0.0152 (0.0045)***
Gain ² x female		0.0014 (0.0018)
Gain ² x high income		0.0033 (0.0018)*
Gain ² x smoker		0.0224 (0.0027)***
Gain ² x healthy life		-0.0062 (0.0028)**
Gain ² x carer		-0.0143 (0.0021)***
Gain ² x LE45		-0.0021 (0.0034)
Gain ² x LE60		-0.0048 (0.0038)
Gain ² x LE75		0.0021 (0.0035)
Lnsig2u	-13.3337 (13.7591)	-13.2098 (16.7375)
Sigma u	0.0013 (0.0088)	0.0014 (0.0113)
Log likelihood	-5567	-5496

Levels of statistical significance: *=10%; **=5%; ***=1%

While the RE probit and RE logit have quite different coefficients, this is predominantly a scale effect, which can be seen by scatter plotting the coefficients. This is done using Utility Function A in Figure 35, and using Utility Function B in Figure 36.

Figure 35: Comparison of Coefficients under Utility Function A

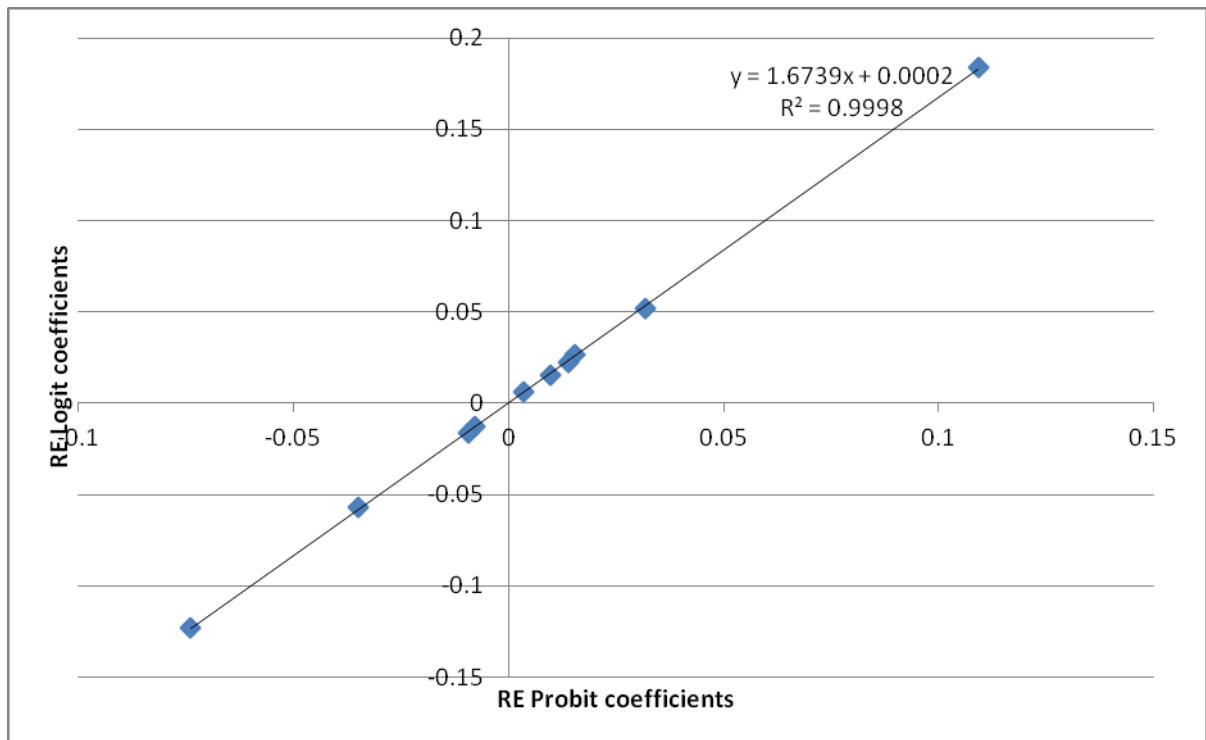
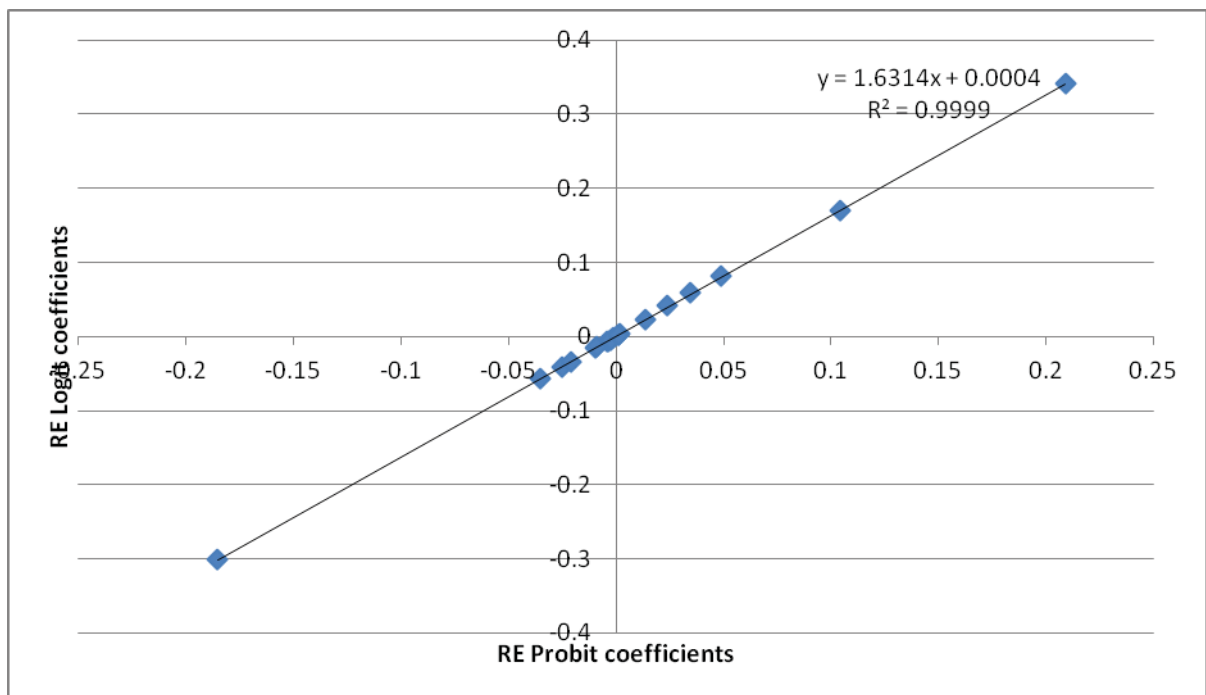


Figure 36: Comparison of Coefficients under Utility Function B



Responses constrained under utility function 1 were willing to discriminate in favour of programmes with a greater health gain, and to recipients who had a lower income, were non-smokers, were carers, or had life expectancies of 45 (relative to those with the base life expectancy of 30 years). The pattern over life expectancy is difficult to explain; the middle

life expectancies appear to be favoured but the cause of this is uncertain. Under utility function 2, similar patterns occur, other than that the discrimination pattern over life expectancy drops out. The quadratic terms are statistically significant at the 5% level for the main effect on *GAIN* (suggesting diminishing marginal utility of time), and on smoking (positive), healthy lifestyles and carer status (both negative). Thus, the discrimination against smokers exhibited throughout is relatively larger for smaller values of *GAIN*, which the discrimination in favour of those with healthy lifestyles or with dependents was relatively larger for smaller values of *GAIN*.

The equity weights produced using these results will be presented following the investigation of heterogeneity.

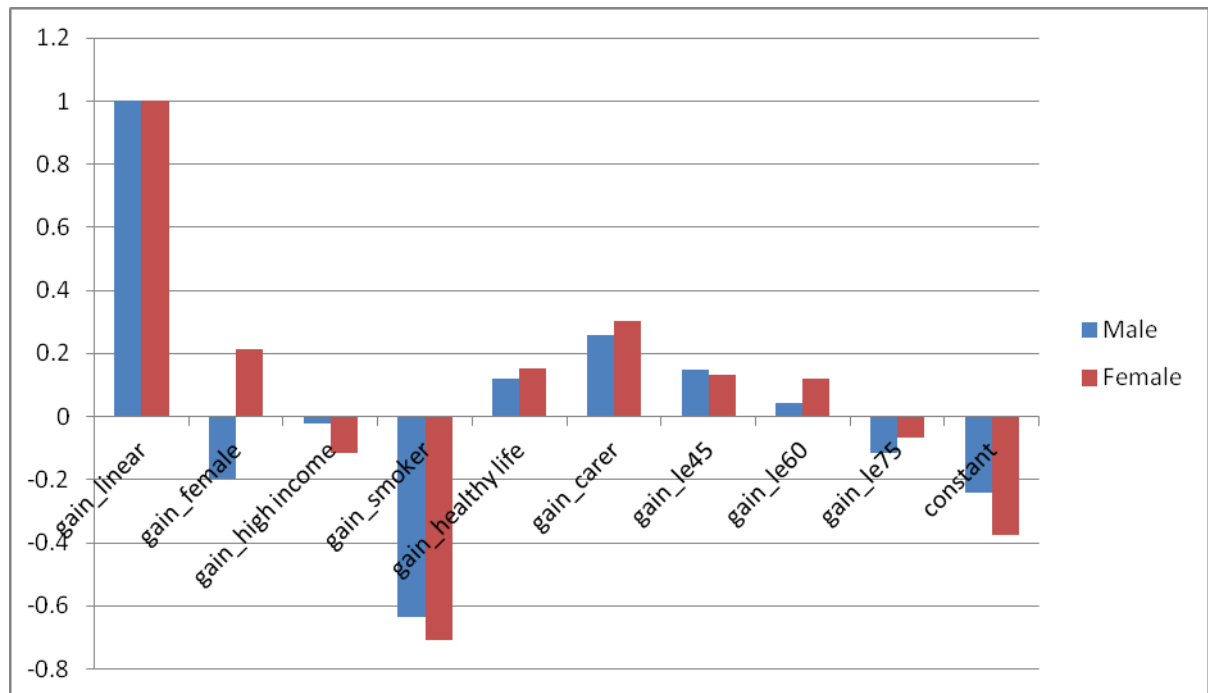
Heterogeneity based on observed respondent characteristics

Gender

The RE probit was repeated investigating the impact of the gender of the survey respondent. The RE probit result in Table 44 was treated as the restricted model (in that it assumes the coefficients for male and female respondents are identical). An unrestricted model was run in which each variable in the RE probit was interacted with a dummy variable equal to 1 if the respondent was female. The likelihood ratio test demonstrated that it was inappropriate to nest the restricted model within the unrestricted one ($p=0.0000$); therefore, gender of the survey respondent influenced the responses they provided.

As proposed in Chapter 3, the coefficients can be compared graphically by rescaling them such that the coefficient on a numeraire was 1. In this case, all coefficients were divided by the coefficient on *GAIN*. The results for the gender subgroup analysis are provided in Figure 37, with the regressions for each gender of respondent provided in Appendix 11.

Figure 37: RE Probit sub-group analysis (gender)

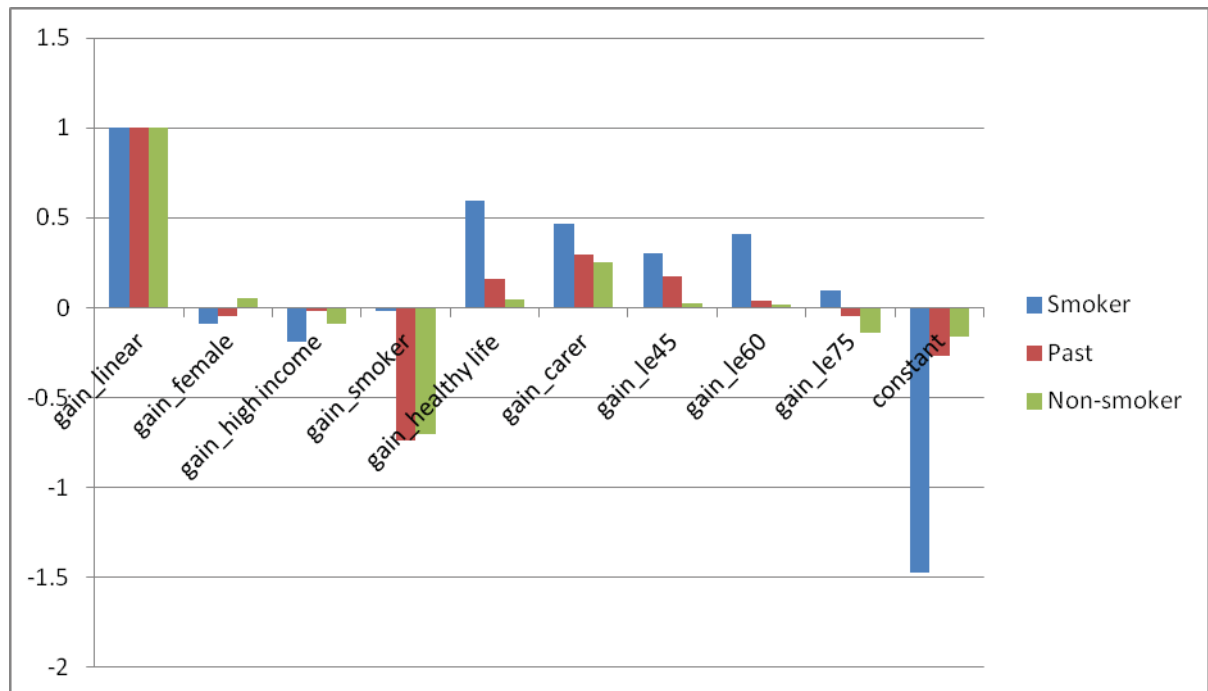


The direction of each of the coefficients is common between genders for each of the dimensions of the experiment with the exception of gender. The average respondent relatively favours health gains accruing in the hypothetical groups with their gender; this is somewhat surprising as the experiment is clearly based on hypothetical groups. It is noteworthy that both coefficients (*GAIN_female* based on male respondents and *GAIN_female* based on female respondents) are statistically significantly different from zero.

Smoking Status

Using the same approach for the smoking status of the respondent, the appropriateness of nesting a restricted model (in which smoking status of the respondent is assumed to be unimportant) within an unrestricted one was tested. As with the case of the gender of the respondent, the LR test rejects the assumption of nesting ($1r \chi^2(9)=141.21; p=0.0000$). The graphical comparison of responses by smoking status considers three sub-groups, namely smokers, former smokers and people who have never been smokers. These results are illustrated in Figure 38.

Figure 38: RE Probit sub-group analysis (smoking)



As with the gender case, the respondents tended to display quite different preferences in the dimension over which the sample was split (in this case, smoking). On average, smokers did not strongly favour either smokers or non-smokers. However, the other two groups strongly discriminate against smokers. Thus, this is further supporting evidence that people, when faced with a hypothetical resource allocation decision, tend to prefer allocating resources to (hypothetical) people with similar characteristics to themselves.

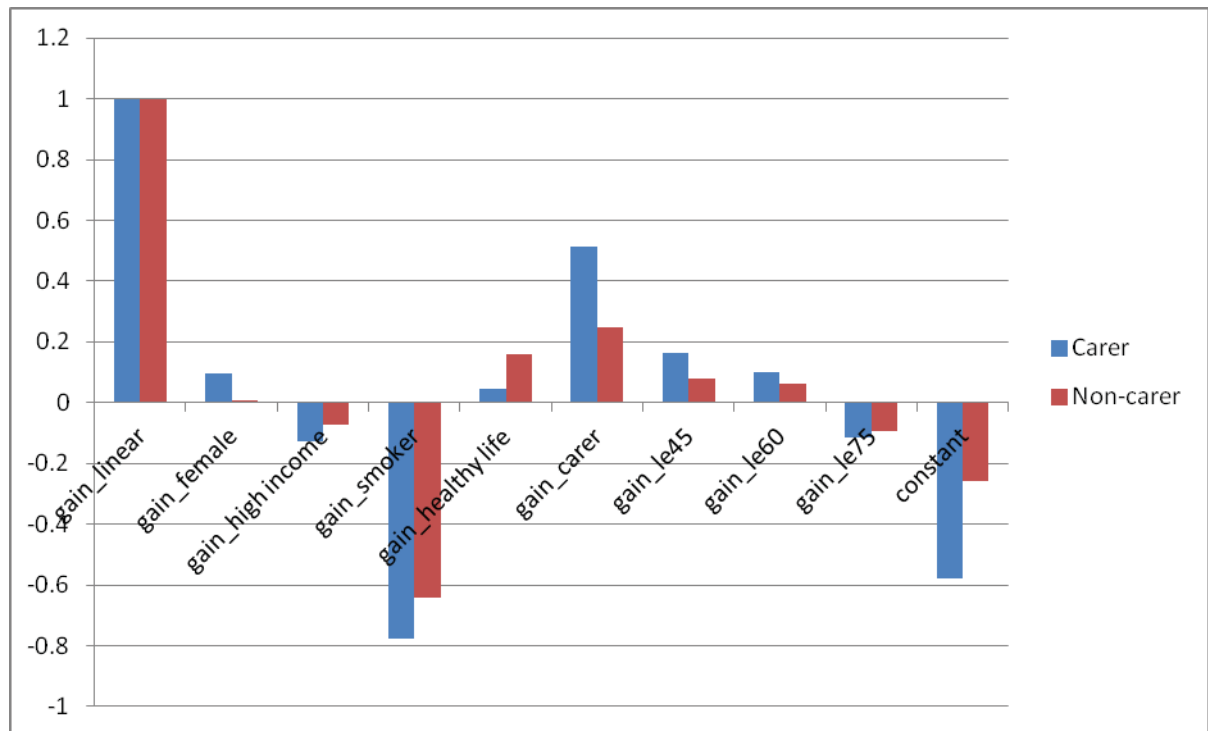
The constant term appears absolutely large, particularly in those who are currently smokers. The negative coefficient means that option A was selected in preference to Option B more than 50%; in this case, the smokers selected A in 52.7% of their choice sets (compared with 51.8% in ex-smokers and 50.8% in those who have never been smokers).

The graphical comparison of responses by smoking status considers three sub-groups, namely smokers, former smokers and people who have never been smokers. These results are illustrated in Figure 39.

Carer Status

Using the same approach for the carer status of the respondent, the appropriateness of nesting a restricted model (in which the carer status of the respondent is assumed to be unimportant) within an unrestricted one was tested. As with the previous two instances, the LR test rejects the assumption of nesting ($1r \chi^2(9)=43.17; p=0.0000$).

Figure 39: RE Probit sub-group analysis (carer status)



As with the previous two sub-group analyses, the log file with the two regressions is presented in Appendix 13.

Again, this suggests that there is a relationship between an individual having a particular attribute and valuing that attribute more highly than other respondents do. As before, there is evidence that carers are likely to value more highly health gains accruing to carers. However, both groups relatively favour gains accruing to carers, reflecting the strong mean response in favour of health gains accruing to carers displayed in Table 44. Thus, the conclusion from these three examples is not that people absolutely discriminate in favour of people like themselves (as was the case in the gender example), but that seeing a characteristic in a hypothetical respondent makes the respondent more likely to favour them.

Modelling heterogeneity

The next step is to consider the various approaches to data modelling described in Chapter 3. This is first in the context of Utility Function A, in which linearity of utility with respect to time is imposed, and then in Utility Function B in which it is not.

Utility function A

Having considered response heterogeneity based on observable characteristics, the next component of the chapter will consider the approaches to modelling heterogeneity more

generally, ranging from the conditional logit to the G-MNL. As in Chapter 5, the results are presented first under Utility Function A (Table 46), then under Utility Function B.

Table 46: Heterogeneity Modelling Results (Utility Function 1)

	Conditional logit			Scale MNL			Unrelated coefficients			Correlated coefficients‡		
	A1	A2	A3	A4	A5#	A6#	Mixed logit	G-MNL	Mixed logit	G-MNL	Mixed logit	G-MNL
Mean												
Model												
Gain	0.184 (0.012)***	0.343 (0.032)***	0.474 (0.032)***	1.484 (0.207)***	0.631 (0.047)***	2.506 (0.525)***	0.474 (0.032)***	1.484 (0.207)***	0.631 (0.047)***	2.506 (0.525)***	0.631 (0.047)***	2.506 (0.525)***
Female	0.006 (0.004)	0.015 (0.005)***	0.021 (0.012)*	0.053 (0.022)**	0.025 (0.015)*	0.045 (0.025)*	0.021 (0.012)*	0.053 (0.022)**	0.025 (0.015)*	0.045 (0.025)*	0.025 (0.015)*	0.045 (0.025)*
High Income	-0.013 (0.005)***	-0.013 (0.006)**	-0.031 (0.011)***	-0.058 (0.024)**	-0.039 (0.014)***	-0.161 (0.037)***	-0.031 (0.011)***	-0.058 (0.024)**	-0.039 (0.014)***	-0.161 (0.037)***	-0.039 (0.014)***	-0.161 (0.037)***
Smoker	-0.123 (0.006)***	-0.229 (0.020)***	-0.290 (0.020)***	-0.991 (0.150)***	-0.367 (0.028)***	-1.284 (0.263)***	-0.290 (0.020)***	-0.991 (0.150)***	-0.367 (0.028)***	-1.284 (0.263)***	-0.367 (0.028)***	-1.284 (0.263)***
Healthy Lifestyle	0.026 (0.008)***	0.058 (0.017)***	0.049 (0.017)***	0.186 (0.046)***	0.077 (0.022)***	0.313 (0.076)***	0.049 (0.017)***	0.186 (0.046)***	0.077 (0.022)***	0.313 (0.076)***	0.077 (0.022)***	0.313 (0.076)***
Carer	0.052 (0.005)***	0.068 (0.007)***	0.142 (0.014)***	0.435 (0.073)***	0.168 (0.017)***	0.437 (0.084)***	0.142 (0.014)***	0.435 (0.073)***	0.168 (0.017)***	0.437 (0.084)***	0.168 (0.017)***	0.437 (0.084)***
Life expect 45	0.023 (0.009)***	0.003 (0.010)	0.042 (0.017)**	0.126 (0.065)*	0.033 (0.026)	0.151 (0.043)***	0.042 (0.017)**	0.126 (0.065)*	0.033 (0.026)	0.151 (0.043)***	0.033 (0.026)	0.151 (0.043)***
Life expect 60	0.016 (0.010)	-0.029 (0.013)**	0.020 (0.021)	0.059 (0.065)	-0.004 (0.032)	0.020 (0.051)	0.020 (0.021)	0.059 (0.065)	-0.004 (0.032)	0.020 (0.051)	-0.004 (0.032)	0.020 (0.051)
Life expect 75	-0.016 (0.009)*	-0.068 (0.014)***	-0.055 (0.025)**	-0.118 (0.054)**	-0.092 (0.037)**	-0.326 (0.079)***	-0.055 (0.025)**	-0.118 (0.054)**	-0.092 (0.037)**	-0.326 (0.079)***	-0.092 (0.037)**	-0.326 (0.079)***
τ		1.247 (0.091)		1.516 (0.114)		1.942 (0.150)***		1.516 (0.114)		1.942 (0.150)***		1.942 (0.150)***
Standard Deviation†												
Gain			0.345 (0.031)***	0.985 (0.134)***	0.701 (0.053)***	1.577 (0.312)***	0.345 (0.031)***	0.985 (0.134)***	0.701 (0.053)***	1.577 (0.312)***	0.345 (0.031)***	1.577 (0.312)***
Female			0.221 (0.015)***	0.701 (0.094)***	0.267 (0.017)***	0.750 (0.150)***	0.221 (0.015)***	0.701 (0.094)***	0.267 (0.017)***	0.750 (0.150)***	0.267 (0.017)***	0.750 (0.150)***
High Income			0.161 (0.015)***	0.481 (0.097)***	0.220 (0.018)***	0.428 (0.085)***	0.161 (0.015)***	0.481 (0.097)***	0.220 (0.018)***	0.428 (0.085)***	0.220 (0.018)***	0.428 (0.085)***
Smoker			0.336 (0.022)***	0.839 (0.113)***	0.479 (0.034)***	0.951 (0.192)***	0.336 (0.022)***	0.839 (0.113)***	0.479 (0.034)***	0.951 (0.192)***	0.479 (0.034)***	0.951 (0.192)***
Healthy Lifestyle			0.173 (0.024)***	0.361 (0.069)***	0.230 (0.030)***	0.592 (0.123)***	0.173 (0.024)***	0.361 (0.069)***	0.230 (0.030)***	0.592 (0.123)***	0.230 (0.030)***	0.592 (0.123)***
Carer			0.228 (0.015)***	0.651 (0.087)***	0.289 (0.019)***	0.877 (0.173)***	0.228 (0.015)***	0.651 (0.087)***	0.289 (0.019)***	0.877 (0.173)***	0.289 (0.019)***	0.877 (0.173)***
Life expect 45			0.093 (0.032)***	0.045 (0.053)	0.326 (0.036)***	0.811 (0.163)***	0.093 (0.032)***	0.045 (0.053)	0.326 (0.036)***	0.811 (0.163)***	0.326 (0.036)***	0.811 (0.163)***
Life expect 60			0.117 (0.039)***	0.275 (0.049)***	0.506 (0.041)***	1.347 (0.258)***	0.117 (0.039)***	0.275 (0.049)***	0.506 (0.041)***	1.347 (0.258)***	0.506 (0.041)***	1.347 (0.258)***
Life expect 75			0.409 (0.032)***	1.114 (0.148)***	0.769 (0.050)***	2.127 (0.426)***	0.409 (0.032)***	1.114 (0.148)***	0.769 (0.050)***	2.127 (0.426)***	0.769 (0.050)***	2.127 (0.426)***
γ				0.034 (0.017)		0.081 (0.022)		0.034 (0.017)		0.081 (0.022)		0.081 (0.022)
Log likelihood	-5570	-5466	-5008	-4938	-4826	-4775	-5008	-4938	-4826	-4775	-4826	-4775
Degrees of freedom	9	10	18	20	54	56	18	20	54	56	54	56
AIC	11158	10952	10052	9916	9760	9662	10052	9916	9760	9662	9760	9662

BIC (n= individuals)	11197	10995	10130	10002	9993	9904
BIC (n = observations)	11222	11030	10192	10072	10180	10097

Statistical significance noted at the 1% level (****), the 5% level (**) and the 10% level (*)

† STATA reports standard deviations as both positive and negative, but notes that the sign is irrelevant. Therefore, the absolute values are presented here

‡ Here, the standard deviations for each of the random variables were generated to allow some comparability with other models generating standard deviation figures. These were generated using the STATA *gmmcov, sd* command. The variance-covariance matrices for Models A5 and A6 are reported in Appendix 14.

The relationships between characteristics of the hypothetical groups and the coefficients in Model A1 are as expected given the patterns identified in Figure 34, and the results obtained through the base case random-effect probit and logit models. Thus, in Model A1, the mean respondent is willing to discriminate in favour of people with low incomes, non-smokers, people with a healthy lifestyle and carers. Once again, the pattern on life expectancy is less clear. Relative to people with a life expectancy of 30, the mean respondent appears to value the health of people with a life expectancy of 45 more highly.

Modelling scale in Model A2 leads to a significant improvement in explanatory power; as noted previously, the one additional parameter appears to be a valuable addition (in terms of Information Criteria). Interestingly, the pattern regarding life expectancy is closer to the a priori assumption (that those with lower life expectancies would be relatively favoured). Moving from Model A1 to Model A3 (i.e. the mixed logit without correlations) also appears sensible, particularly as all standard deviations are statistically significant at the 1% level. Combining Models A2 and A3 in Model A4 (estimating a G-MNL without correlation) outperforms all three less flexible models, with AIC and BIC improving again. The γ term is low suggesting that G-MNL-II (which was described in Chapter 3) is the more appropriate specification in this case. This is a similar finding to that in Chapter 5. However, it should be noted that the change in log likelihood from A3 to A4 is smaller than that between A1 and A2, even though the former move involves the estimation of both τ and γ , rather than simply τ . It might be concluded that the explanatory benefit of adding a parameter is inversely related to the flexibility of the base case model, simply because there is less variability to explain.

The benefit of allowing for correlations in Models A5 and A6 is uncertain based on Information Criteria. The impact on log-likelihood (182 points in the mixed logit and 163 in the G-MNL) appears worthwhile using either the AIC or BIC (if the latter is based on the number of survey respondents). However, if the n term is based on the number of observations (which is the default in the STATA command), the benefit of allowing correlations is marginal (in that it leads to a slight improvement in the mixed logit case, but a slight deterioration in the G-MNL).

Utility function B

The corresponding results under the non-linear utility function are presented in Table 47.

Table 47: Heterogeneity Modelling Results (Non-Linear Utility Function)

Mean (s.e.) Model	Conditional logit		Scale MNL		Uncorrelated coefficients		Correlated coefficients‡	
	B1		B2		Mixed logit	G-MNL	Mixed logit	G-MNL
Gain	0.341 (0.047)***		0.437 (0.064)***		0.550 (0.072)***	1.208 (0.204)***	0.627 (0.086)***	2.028 (0.426)***
Gain ²	-0.015 (0.005)***		-0.016 (0.006)***		-0.010 (0.007)	0.015 (0.017)	0.000 (0.008)	0.030 (0.020)
Female	-0.006 (0.016)		0.014 (0.018)		-0.007 (0.024)	0.008 (0.048)	0.000 (0.028)	0.021 (0.061)
High Income	-0.042 (0.017)**		-0.034 (0.019)*		-0.052 (0.025)**	-0.015 (0.048)	-0.066 (0.028)**	-0.223 (0.082)***
Smoker	-0.300 (0.022)***		-0.430 (0.037)***		-0.424 (0.036)***	-1.000 (0.125)***	-0.490 (0.042)***	-1.813 (0.318)***
Healthy Lifestyle	0.082 (0.027)***		0.105 (0.037)***		0.106 (0.040)***	0.253 (0.085)***	0.147 (0.046)***	0.461 (0.140)***
Carer	0.170 (0.018)***		0.202 (0.022)***		0.242 (0.028)***	0.616 (0.084)***	0.270 (0.032)***	1.021 (0.189)***
Life expect 45	0.040 (0.032)		0.034 (0.034)		0.087 (0.046)*	0.234 (0.078)***	0.097 (0.055)*	0.322 (0.143)**
Life expect 60	0.058 (0.037)		0.036 (0.041)		0.106 (0.054)**	0.220 (0.104)**	0.127 (0.064)**	0.067 (0.153)
Life expect 75	-0.035 (0.033)		-0.110 (0.039)***		-0.088 (0.052)*	-0.089 (0.109)	-0.079 (0.064)	-0.444 (0.184)**
Gain x Female	0.001 (0.002)		-0.000 (0.002)		0.004 (0.003)	0.005 (0.006)	0.004 (0.003)	0.001 (0.008)
Gain x High Income	0.003 (0.002)*		0.002 (0.002)		0.003 (0.003)	-0.006 (0.006)	0.004 (0.003)	0.002 (0.008)
Gain x Smoker	0.022 (0.003)***		0.031 (0.004)***		0.020 (0.004)***	0.024 (0.009)***	0.019 (0.005)***	0.065 (0.015)***
Gain x Healthy Lifestyle	-0.006 (0.003)**		-0.007 (0.004)*		-0.007 (0.005)	-0.013 (0.010)	-0.009 (0.005)*	-0.029 (0.014)**
Gain x Carer	-0.014 (0.002)***		-0.016 (0.002)***		-0.015 (0.003)***	-0.037 (0.007)***	-0.015 (0.004)***	-0.055 (0.013)***
Gain x LE45	-0.002 (0.003)		-0.004 (0.004)		-0.006 (0.005)	-0.016 (0.008)*	-0.009 (0.007)	-0.013 (0.015)
Gain x LE60	-0.005 (0.004)		-0.006 (0.004)		-0.011 (0.006)*	-0.025 (0.011)**	-0.018 (0.007)**	-0.009 (0.018)
Gain x LE75	0.002 (0.004)		0.006 (0.004)		0.005 (0.006)	0.008 (0.011)	-0.002 (0.007)	0.016 (0.018)
τ			1.050 (0.078)			1.407 (0.096)***		1.674 (0.119)
Standard Deviation †								
Gain					0.325 (0.031)***	0.910 (0.108)***	0.678 (0.054)***	2.426 (0.434)***
Female					0.208 (0.014)***	0.562 (0.068)***	0.262 (0.018)***	1.013 (0.178)***
High Income					0.151 (0.015)***	0.357 (0.043)***	0.213 (0.019)***	0.450 (0.078)***
Smoker					0.319 (0.021)***	0.732 (0.090)***	0.462 (0.033)***	1.884 (0.334)***

Healthy Lifestyle			0.161 (0.024)***	0.292 (0.048)***	0.216 (0.030)***	0.775 (0.151)***
Caret			0.216 (0.015)***	0.551 (0.066)***	0.278 (0.019)***	0.927 (0.165)***
Life expect 45			0.069 (0.035)**	0.012 (0.050)	0.326 (0.037)***	0.989 (0.182)***
Life expect 60			0.094 (0.042)**	0.170 (0.039)***	0.500 (0.041)***	1.778 (0.312)***
Life expect 75			0.388 (0.031)***	0.891 (0.111)***	0.756 (0.050)***	2.763 (0.484)***
γ				0.032 (0.019)		0.000
Log likelihood	-5499	-5405	-4981	-4919	-4805	-4739
Degrees of freedom	18	19	27	29	63	64
AIC	11034	10848	10016	9897	9735	9606
BIC (n= individuals)	11174	10930	10133	10021	10008	9882
BIC (n = observations)	11112	10996	10226	10122	10226	10104

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

† STATA reports standard deviations as both positive and negative, but notes that the sign is irrelevant. Therefore, the absolute values are presented here

‡ Here, the standard deviations for each of the random variables were generated to allow some comparability with other models generating standard deviation figures. These were generated using the STATA *gmmcov, sd* command. The variance-covariance matrices for Models B5 and B6 are reported in Appendix 15.

See text below for details regarding estimation of Models B5 and B6

As Gu *et al.* (2011) note, the convergence of the G-MNL is highly dependent on the specification of starting values. Models B1-B4 did not require specification of starting values to elicit convergence. However, Models B5 and B6 either did not converge, or converged poorly (in the sense that they were outperformed in log-likelihood terms by their corresponding Utility Function 1 models, which are nested within them). Initially, B5 was estimated without specifying starting values, which led to a poorer log likelihood than Model A5 (-4851), despite A5 being nested in B5. Therefore, the coefficients from A5 were used as starting values for a re-estimation of Model B5, with the coefficients on the additional higher-order coefficients starting at 0. For model B6, the coefficients on everything other than γ and τ from model B5 were used, although they were multiplied by 4 to approximate the increase in coefficient size noted between Models A5 and A6. The starting values for γ and τ were initially taken from model A6, and allowed to be freely estimated. However, this failed to converge. Therefore, a modified approach was taken in which the value of γ was taken from Model A6 (0.081) and fixed at that point. The remaining coefficients were then estimated with this additional constraint imposed. Again, this failed to converge. The approach which did converge was to set γ to be fixed at 0. This imposes G-MNL-II on the data.

Allowing for non-linearity of utility with respect to *GAIN* appears to improve model fit as the coefficient on the quadratic term is statistically significant (and negative, which is a conventional finding suggesting diminishing utility of extra time). The pattern of coefficients in the interaction terms in Model B1 has four statistically significant coefficients out of eight, two of them at the 1% level. The impact of relaxing the utility function to allow the interactions between the quadratic term and the characteristics of the hypothetical population are difficult to interpret. The coefficient on $\text{Gain}^2 \times \text{Smoker}$ is positive and statistically significant, suggesting the discrimination against smokers in the main effect term is larger for relatively short gains: the opposite is true for the term on carer status (i.e. people discriminate more in favour of carers when gains are short).

Relative to Model A1, Model B1 appears to perform well. However, comparing it with Model A3, which has the same number of degrees of freedom (18), it performs poorly suggesting that, in this instance, allowing for preference heterogeneity is a more productive way of explaining responses than allowing for non-linearity.

As under Utility Function 1, allowing for scale performs well (B2), as does allowing for preference heterogeneity (B3), and for both (B4), as the Information Criteria reduce monotonically from B1 to B4. However, it is instructive to look at the point improvement from allowing non-linearity in the approaches B1-B4. In B1, the log likelihood improves by 71 points. This reduces to 61 in B2, 27 in B3, and 16 in B4. In terms of AIC, Utility Function 2 is preferred to Utility Function 1 across 1-4. However, in terms of the BIC where n is defined as the number of individuals, A3 and B3 are considered equally good. However, if n is assumed to be the number of observations, A3 is actually preferred to B3. This raises the question about whether it is worthwhile to allow for both a non-linear utility function over time and for response heterogeneity.

Allowing for correlations in Models B5 and B6 appears unwarranted based on Information Criteria. While the correlation matrices for these two models contain a proportion of statistically significant terms, the impact on log-likelihood does not appear to produce a large enough improvement to justify the proliferation of estimated parameters. Additionally, it is apparent that Models B5 and B6 are outperformed by Models A5 and A6 respectively. It should be noted that Model A6 is not strictly nested within Model B6 as the latter failed to converge with a freely estimated γ . However, I would argue that constraining γ at the value derived from Model A6 is unlikely to represent a large constraint on the analysis.

Model comparison

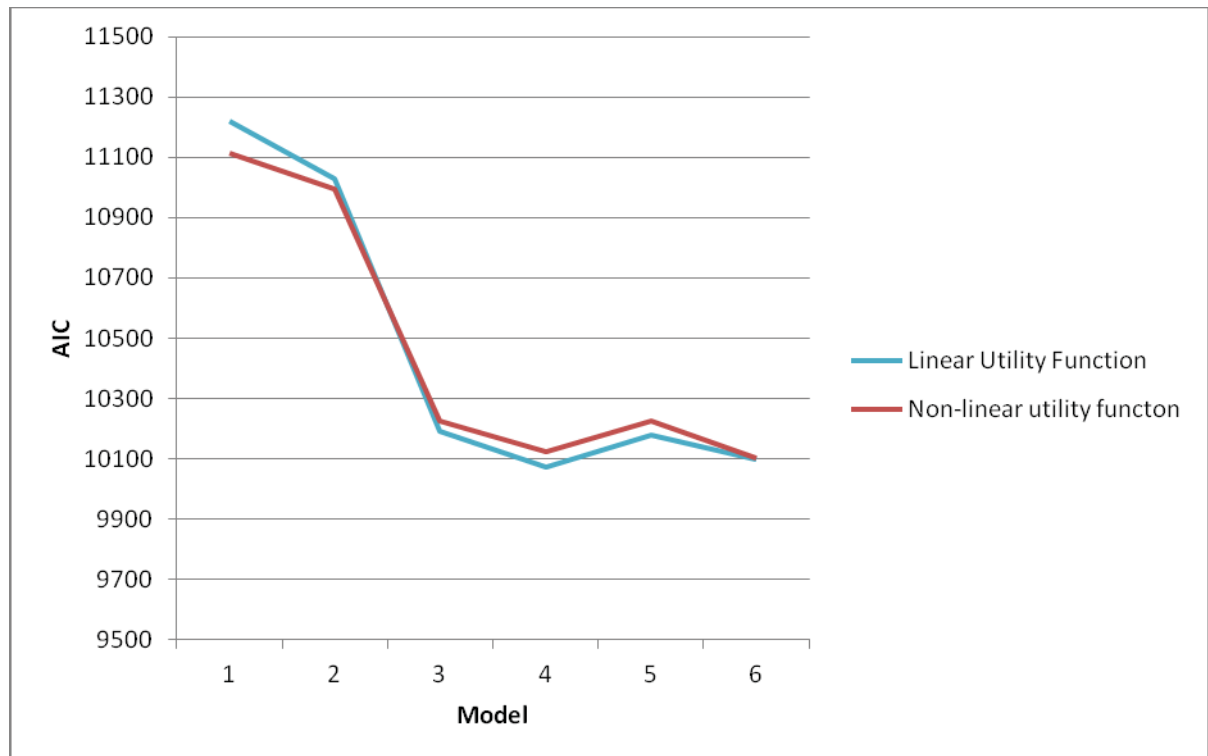
The summary of AIC and BIC, and the ranking under each of the twelve are presented in Table 34.

Table 48: Model Comparison

Model	AIC	BIC (n=individuals)	BIC (n=observations)
A1	11158	11197	11222
A2	10952	10995	11030
A3	10052	10130	10192
A4	9916	10002	10072
A5	9760	9993	10180
A6	9662	9904	10097
B1	11034	11174	11112
B2	10848	10930	10996
B3	10016	10133	10226
B4	9897	10021	10122
B5	9735	10008	10226
B6	9606	9882	10104

The pattern on Information Criteria is different to that demonstrated in the previous chapter. The three criteria are graphed across models and utility functions in Figure 40, Figure 41 and Figure 42.

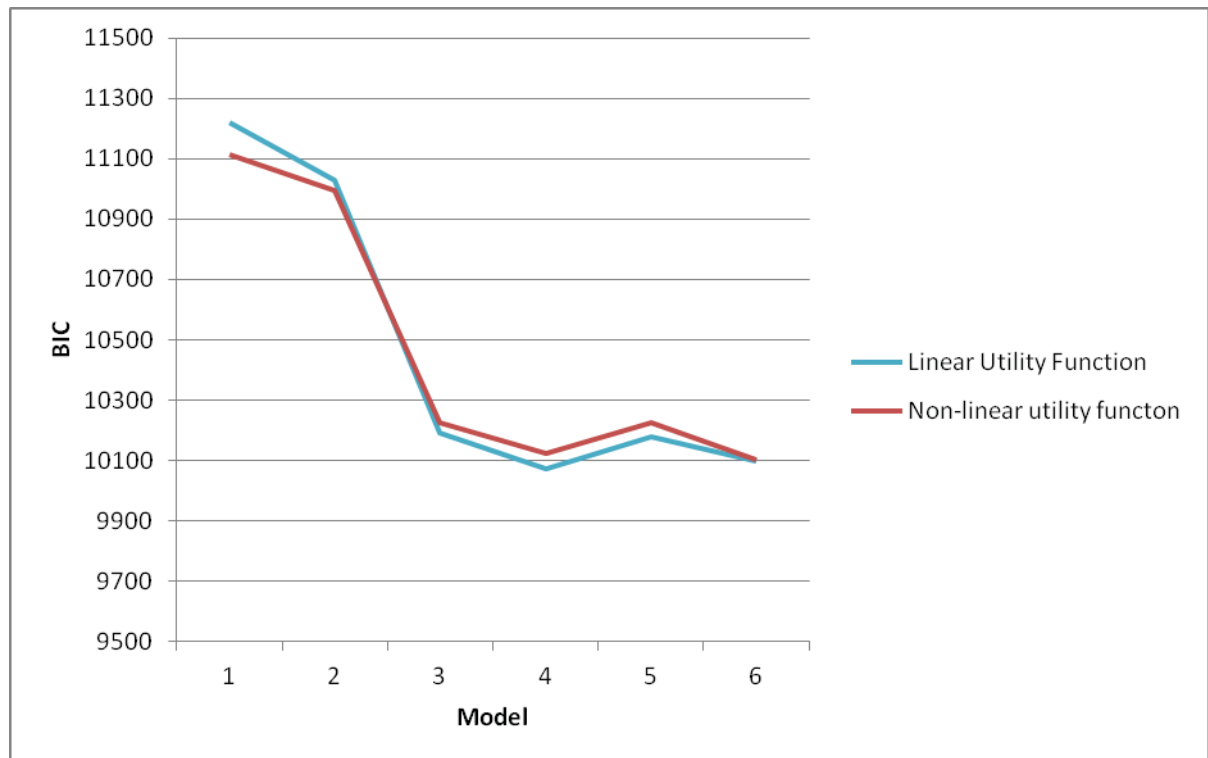
Figure 40: AIC figures for the 12 Models



Under the AIC, both utility functions suggest that the less constrained models perform better. That is, the AIC reduces monotonically as the number of degrees of freedom increases. As with the SF-6D DCE results in Chapter 5, the impact on AIC of modelling scale is large, reflecting both the improved model fit and parsimony of the approach. The main difference between the results in Chapter 5 and here concern the merits of constraining utility to be linear with respect to time. In Chapter 5, the non-linear model outperformed the linear one. However the pattern is less clear here. Under Models A1-A2, the non-linear model performs better. However, in all other models, the impact of allowing a non-linear utility function is very small.

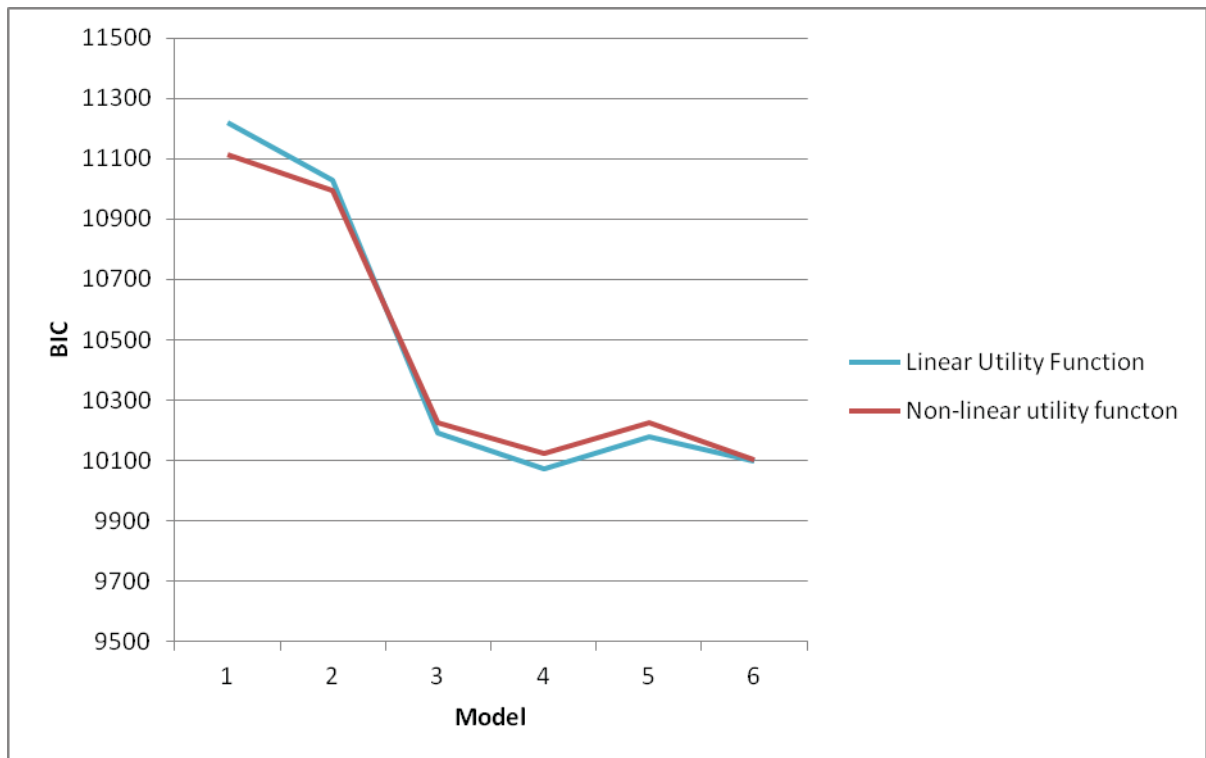
The corresponding figures for the BIC, where n is the number of individuals (rather than the number of observations) is presented in Figure 41.

Figure 41: BIC figures for the 12 Models (n=individuals)



If the BIC is employed, the pattern is similar. The move to Utility Function B (reflecting non-linearity of utility) is justified under 1 and 2, but 3-6 suggest it to be of marginal use. The model fit results from the alternative BIC measure (where n reflects the number of observations) is described in Figure 42.

Figure 42: BIC figures for the 12 Models (n=observations)



If the n term is used to determine BIC is taken to be the number of observations, the patterns observed in the alternative BIC become more pronounced. In this case, monotonicity breaks down in both Utility Function A and B, with the uncorrelated G-MNL emerging as the best model in Utility Function A. The merits of moving to a non-linear utility function are even more marginal under this BIC. While B1 and B2 outperforms A1 and A2 respectively, 3-6 all favour using the more restricted approach.

Generating equity weights

To this point, I have identified that people are willing to apply differential valuation to outputs accruing to different people. The next stage is to convert these results into something which aims to be useful in health policy decisions. The equity weights derived under the methods proposed in Chapter 3, and using the CV approach suggested by Lancsar and Savage (2004) are presented in Table 49. Confidence intervals are generated for each of the equity weights using a bootstrapping approach with 50 replications. To allow comparison with the CV-derived weights, those CV- derived weights that fall outside the confidence interval are highlighted in **bold**.

Table 49: Equity Weights

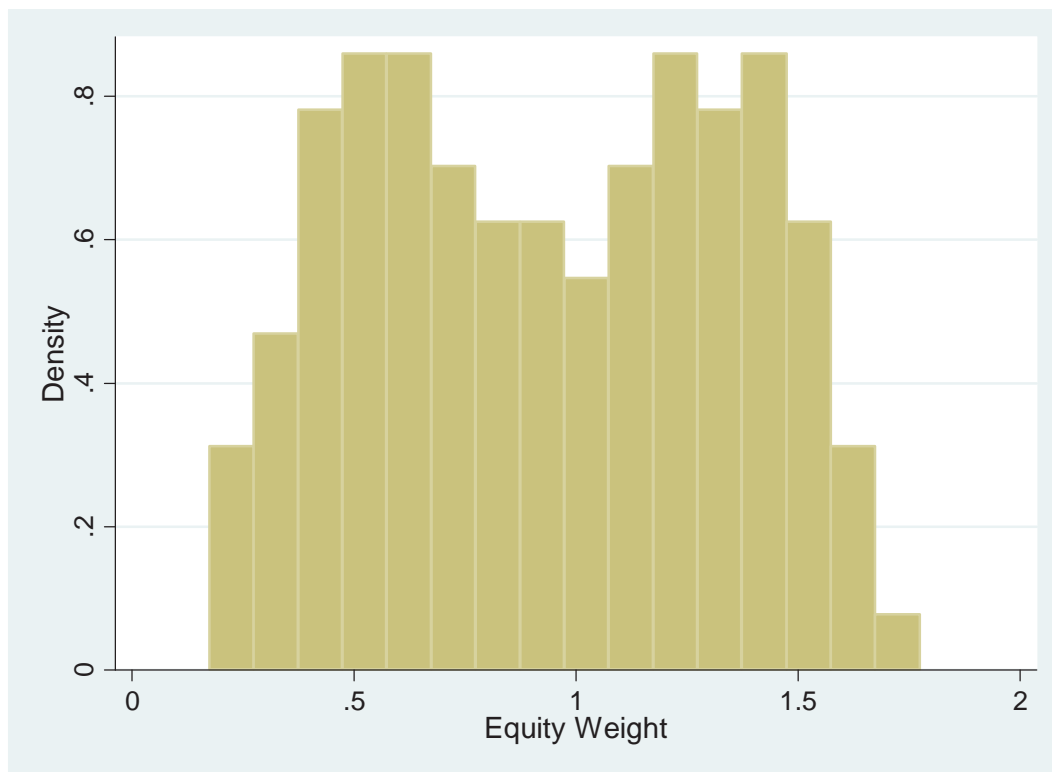
Income	Smoker	Healthy life?	Carer?	Life expectancy	Male Equity Weight (95% CI)	CV Weight	Female Equity Weight (95% CI)	CV Weight
High	Yes	Yes	Yes	30	0.72 (0.58-0.86)	0.817	0.75 (0.62-0.89)	0.850
High	Yes	Yes	Yes	45	0.86 (0.71-1.00)	0.945	0.89 (0.73-1.05)	0.978
High	Yes	Yes	Yes	60	0.81 (0.71-0.92)	0.906	0.85 (0.72-0.98)	0.939
High	Yes	Yes	Yes	75	0.63 (0.54-0.71)	0.731	0.66 (0.59-0.74)	0.763
High	Yes	Yes	No	30	0.41 (0.30-0.53)	0.527	0.45 (0.32-0.58)	0.559
High	Yes	Yes	No	45	0.55 (0.43-0.67)	0.655	0.58 (0.45-0.72)	0.687
High	Yes	Yes	No	60	0.51 (0.38-0.64)	0.616	0.54 (0.44-0.64)	0.648
High	Yes	Yes	No	75	0.32 (0.23-0.42)	0.441	0.36 (0.26-0.45)	0.473
High	Yes	No	Yes	30	0.57 (0.44-0.70)	0.677	0.61(0.49-0.72)	0.709
High	Yes	No	Yes	45	0.71 (0.57-0.85)	0.804	0.74 (0.63-0.85)	0.837
High	Yes	No	Yes	60	0.67 (0.55-0.78)	0.765	0.70 (0.60-0.80)	0.798
High	Yes	No	Yes	75	0.48 (0.39-0.57)	0.590	0.51 (0.43-0.60)	0.622
High	Yes	No	No	30	0.27 (0.15-0.38)	0.386	0.30 (0.16-0.44)	0.419
High	Yes	No	No	45	0.40 (0.28-0.52)	0.514	0.43 (0.28-0.59)	0.546
High	Yes	No	No	60	0.36 (0.26-0.46)	0.475	0.39 (0.28-0.51)	0.507
High	Yes	No	No	75	0.17 (0.05-0.30)	0.300	0.21 (0.07-0.35)	0.332
High	No	Yes	Yes	30	1.43 (1.27-1.60)	1.494	1.47 (1.27-1.66)	1.526
High	No	Yes	Yes	45	1.57 (1.39-1.75)	1.622	1.60 (1.42-1.78)	1.654
High	No	Yes	Yes	60	1.53 (1.35-1.71)	1.583	1.56 (1.40-1.72)	1.615
High	No	Yes	Yes	75	1.34 (1.25-1.43)	1.408	1.38 (1.32-1.44)	1.440
High	No	Yes	No	30	1.13 (0.99-1.26)	1.204	1.16 (1.05-1.28)	1.236
High	No	Yes	No	45	1.26 (1.13-1.40)	1.332	1.30 (1.12-1.47)	1.364
High	No	Yes	No	60	1.22 (1.09-1.35)	1.293	1.26 (1.11-1.40)	1.325
High	No	Yes	No	75	1.04 (0.98-1.09)	1.117	1.07 (1.01-1.14)	1.150
High	No	No	Yes	30	1.29 (1.16-1.42)	1.353	1.32 (1.19-1.45)	1.386
High	No	No	Yes	45	1.42 (1.27-1.57)	1.481	1.46 (1.30-1.62)	1.513
High	No	No	Yes	60	1.38 (1.23-1.53)	1.442	1.41 (1.25-1.56)	1.475
High	No	No	Yes	75	1.20 (1.13-1.26)	1.267	1.23 (1.17-1.29)	1.299

High	No	No	No	No	30	0.98 (0.87-1.09)	1.063	1.01 (0.88-1.15)	1.095
High	No	No	No	No	45	1.12 (1.00-1.23)	1.191	1.15 (1.02-1.28)	1.223
High	No	No	No	No	60	1.07 (0.93-1.22)	1.152	1.11 (0.97-1.24)	1.184
High	No	No	No	No	75	0.89 (0.82-0.96)	0.976	0.92 (0.85-1.00)	1.009
Low	Yes	Yes	Yes	Yes	30	0.80 (0.67-0.93)	0.889	0.83 (0.70-0.96)	0.922
Low	Yes	Yes	Yes	Yes	45	0.93 (0.79-1.07)	1.017	0.97 (0.82-1.11)	1.050
Low	Yes	Yes	Yes	Yes	60	0.89 (0.77-1.01)	0.978	0.92 (0.77-1.07)	1.011
Low	Yes	Yes	Yes	Yes	75	0.71 (0.63-0.78)	0.803	0.74 (0.65-0.83)	0.835
Low	Yes	Yes	Yes	No	30	0.49 (0.37-0.61)	0.599	0.52 (0.42-0.63)	0.631
Low	Yes	Yes	Yes	No	45	0.63 (0.51-0.74)	0.727	0.66 (0.55-0.77)	0.759
Low	Yes	Yes	Yes	No	60	0.58 (0.47-0.69)	0.688	0.62 (0.49-0.75)	0.720
Low	Yes	Yes	Yes	No	75	0.40 (0.30-0.50)	0.513	0.43 (0.33-0.53)	0.545
Low	Yes	No	No	Yes	30	0.65 (0.52-0.78)	0.749	0.68 (0.57-0.80)	0.781
Low	Yes	No	No	Yes	45	0.78 (0.64-0.92)	0.876	0.82 (0.69-0.94)	0.909
Low	Yes	No	No	Yes	60	0.74 (0.64-0.85)	0.838	0.78 (0.67-0.88)	0.870
Low	Yes	No	No	Yes	75	0.56 (0.46-0.65)	0.662	0.59 (0.50-0.69)	0.695
Low	Yes	No	No	No	30	0.34 (0.23-0.45)	0.458	0.38 (0.25-0.50)	0.491
Low	Yes	No	No	No	45	0.48 (0.35-0.60)	0.586	0.51 (0.39-0.64)	0.618
Low	Yes	No	No	No	60	0.44 (0.32-0.55)	0.547	0.47 (0.36-0.58)	0.580
Low	Yes	No	No	No	75	0.25 (0.11-0.39)	0.372	0.28 (0.19-0.38)	0.404
Low	No	Yes	Yes	Yes	30	1.51 (1.35-1.67)	1.566	1.55 (1.39-1.70)	1.598
Low	No	Yes	Yes	Yes	45	1.65 (1.43-1.86)	1.694	1.68 (1.50-1.86)	1.726
Low	No	Yes	Yes	Yes	60	1.61 (1.43-1.78)	1.655	1.64 (1.49-1.78)	1.687
Low	No	Yes	Yes	Yes	75	1.42 (1.34-1.50)	1.480	1.45 (1.37-1.53)	1.512
Low	No	Yes	Yes	No	30	1.20 (1.08-1.33)	1.276	1.24 (1.11-1.37)	1.308
Low	No	Yes	Yes	No	45	1.34 (1.17-1.51)	1.404	1.37 (1.20-1.55)	1.436
Low	No	Yes	Yes	No	60	1.30 (1.18-1.42)	1.365	1.33 (1.19-1.47)	1.397
Low	No	Yes	Yes	No	75	1.11 (1.05-1.17)	1.189	1.15 (1.08-1.21)	1.222
Low	No	No	No	Yes	30	1.36 (1.19-1.53)	1.425	1.40 (1.26-1.53)	1.458
Low	No	No	No	Yes	45	1.50 (1.34-1.66)	1.553	1.53 (1.35-1.72)	1.585
Low	No	No	No	Yes	60	1.46 (1.32-1.59)	1.514	1.49 (1.32-1.66)	1.547

Low	No	No	Yes	75	1.27 (1.21-1.33)	1.339	1.31 (1.23-1.38)	1.371
Low	No	No	No	30	1.06 (0.95-1.16)	1.135	1.09 (0.98-1.20)	1.167
Low	No	No	No	45	1.19 (1.06-1.33)	1.263	1.23 (1.07-1.38)	1.295
Low	No	No	No	60	1.15 (1.02-1.27)	1.224	1.18 (1.07-1.30)	1.256
Low	No	No	No	75	0.97 (0.89-1.03)	1.048	1.00 (0.94-1.06)	1.081

As expected, the equity weights follow the patterns demonstrated by the RE probit results in Table 44. Thus, the equity weights show that health gains accruing to carers, non-smokers, and people with a healthy lifestyle are relatively favoured. The distribution of equity weights across the 128 hypothetical groups is presented in Figure 43.

Figure 43: Distribution of Equity Weights



These can, in principle, be used in QALY calculations in much the same way as utility weights are used to define the health-related quality of life. Thus, the QALY gains and losses applying to particular groups in society can be weighted up or down (through multiplication of the QALYs and the relevant equity weight) dependent on the characteristics considered in the experiment.

Conclusions and implications

The use of a discrete choice experiment to elicit preferences regarding the allocation of health gains has been shown to be valid, and to be able to produce results which might be adapted for use in economic evaluation. The work identifies that, for the mean respondent, there is a preference for improving health outcomes in those not likely to receive Williams' Fair Innings (1997), in carers, non-smokers, those who lead a healthy life, and those with a

low income. These patterns were consistent across utility function specification, and approach towards the modelling of heterogeneity.

This conclusion contrasts with that of Lancsar *et al.* (2011) who argue that weighting QALYs is generally not appropriate, and would be unlikely to significantly impact on the scale of the gain accruing from a healthcare intervention. The weights presented in this study suggest this conclusion does not hold in our data. Thus, when Lancsar *et al.* (2011) argue that

“(T)he size of the gain dominated the characteristics of the recipients of those gains, suggesting a desire to maximise health and a reluctance to trade off health gain for other characteristics as the health gain increased... The important point is that these results comply with the no-weighting position currently adopted by HTA agencies and governments around the world” (p.475),

the divergence between this finding and my results requires exploration. There are four possible explanations for this divergence. Firstly, it might be that our respondents held different views to those studied by Lancsar *et al.* (2011). However, this is unlikely to be a strong driver of the difference as Australia and the UK are generally considered to be similar culturally. Secondly, the way the question was posed may drive the result. Inadvertent emphasis of certain aspects of the choice may cause results in different experiments to differ. Thirdly, the two studies consider quite different dimensions and, in dimensions that are common to both experiments (such as life expectancy), the levels were different. Lancsar *et al.* (2011) investigated quality of life of hypothetical healthcare recipients and age of onset, which were not explored in the work described in this thesis). Rather, the work in this thesis included a series of characteristics which might impact on the desire of the individual to receive health gain (smoking, carer status, healthy lifestyle). It might be argued that the dimensions selected by Lancsar *et al.* (2011) are ones over which preferences are not strong (and likely to over-ride the conventional maximisation of QALYs). The fourth explanation for the divergence between studies is that the method for converting regression results into QALY weights differs, specifically in that we have employed a marginal willingness to pay approach (MWTP), while Lancsar *et al.* (2011) have adopted the compensating variation (CV). A debate regarding the merits of the two approaches has taken place (Lancsar and Savage, 2004; Ryan, 2004; Santos Silva, 2004); however, both techniques are in current use, so no consensus has yet emerged regarding their relative merits. To investigate the sensitivity of the result to the choice between MWTP and CV, the CV weights are presented in

Appendix 1 alongside the base case results. The values under the CV are closer to 1 than under the MWTP, suggesting that this explanation may contribute to the divergence between the two studies.

One interesting caveat to the results presented in this thesis is that there was a tendency for people to relatively favour health gains that impact on people with similar health profiles to themselves; a result which is somewhat surprising given the task was clearly hypothetical. An important distinction to make in this area is that some of the characteristics over which this pattern is demonstrated are the consequence of a choice of the hypothetical respondents; for example, someone who makes the decision not to be a smoker may feel that someone who has made the same decision may be more deserving of health gain. However, this pattern of discrimination also applies to gender, something not determined by a choice.

The choice of both the RE probit and the simpler utility function 1 as the base case result requires discussion. The more relaxed consideration of heterogeneity typified by the G-MNL and the mixed logit was shown to improve model fit considerably. Similarly, the relaxation of the linearity of the utility function with respect to time does likewise. With regard to the use of the G-MNL or the mixed logit, it should be noted that the impact of doing so on the mean respondent (which is the primary focus of this kind of analysis) is limited (Greene and Hensher, 2011). While we might be interested in segmenting the population into types of respondents, the use of these kinds of results in public policy making is unclear at best. With regard to the use of utility function 1, a judgement is required whether the extra predictable ability of using the more flexible utility function 2 is worth the implication that weights on QALYs are dependent on the value of the GAIN attribute. In this thesis, it was decided that the more relaxed utility function was not appropriate; however, the result that people do not display linearity in this regard is potentially important in certain settings.

One issue which needs to be addressed is the fundamental objection to weighting outcomes in economic evaluation of health technologies. The response to this is to argue that we implicitly weight outcomes anyway, only that these weights are constrained to be equal to one for all groups. The objection is then whether it is ethically defensible to weight outcomes differently for different groups. The results presented here have to overcome two obstacles before they can be reasonably used in practice. First, it has to be shown that respondents accept the resource allocation decisions they follow from the equity weights generated here.

If we accept the respondents are comfortable with the weightings generated, the second obstacle to overcome is the ethical defensibility of the stated preferences. Against what standard should the ethical defensibility of societal preferences be judged? In the introductory chapter, I discussed Empirical Ethics, a non-welfarist approach in which surveys are used to value health, but ethical concerns are used to constrain those revealed preferences. In that discussion, I argued that identifying a societal preference which would be considered unethical is highly unlikely as the methods for determining ethics are uncertain and require a degree of consensus. Since society is unlikely to agree on ethical constraints which counter their own stated preferences, I argued that the constraint was likely to be empty. This applies in this context.

Even if the results presented here are accepted as valid representations of societal preferences, and it is appropriate that preferences of equity be included in the decision making process, it remains uncertain whether a formal quantifiable approach to the inclusion of equity is the best approach. It might be argued that the consideration of equity is necessarily flexible to different settings, which involve finer distinctions than those made in the work presented in this chapter. Flexibility with regard to the consideration of equity issues should be allowed, but the weights derived here should be considered as a guide for decision makers. Thus, decision makers should be aware that society has these patterns of views regarding equity, but that they be allowed to integrate them at their discretion. However, the decisions of these policy makers should be disseminated where possible to allow society to judge whether their decisions are acceptable, and their consideration of equity concerns was justified and well-considered.

Table 50: Equity-Efficiency trade-off search strategy

#	Searches	Hits
1	Welfaris\$ and (Extra-welfaris\$ or Extrawelfaris\$).mp	15
2	Fair Innings.mp	45
3	Egalitarian\$.mp	901
4	Social Welfare Function.mp	35
5	Rule of rescue.mp	41
6	Efficiency-Equity trade-off.mp	0
7	(age adj weight).ab,ti	6349
8	Equity.sh,ab,ti	7698
9	Or/1-7	7380
10	8 and 9	65
11	Remove duplicates from 10	43
12	Limit 11 to english language	34
13	From 12 keep ...	26

mp=ti,ot,ab,nm,hw,sh,tn,dm,mf

Chapter 7: Conclusions and Implications

This thesis has investigated how health outcomes are measured and valued for use in economic evaluation. Economic evaluation has become standard practice in reimbursement decisions across the developed world, so the techniques we use, and the assumptions we make in doing so, play a major role in how scarce resources are allocated. The thesis has illustrated some potential flaws in how orthodox economic evaluation is undertaken, and presented a series of case studies in which areas of weakness are addressed. Importantly, when the current orthodox approach to valuing health outcomes for economic evaluation has been shown to be a simplification or distortion of how society wishes resources to be allocated, the thesis has provided results which can be directly applied to these estimates.

The state of economic evaluation of healthcare

In Chapter 1, I investigated the meaning of, and need for, economic evaluation in healthcare, and how it differs from standard welfare economics orthodoxy. The reason for doing this was to identify the framework within which the valuation of health outcomes typically occurs. The healthcare sector is characterised by endemic market failure (Arrow, 1963), and the dominant position of government as purchaser in most industrialised nations necessitates a framework within which the multitude of possible resource uses can be ranked. Regarding the two major competing theories of valuing health outcomes (welfarism and extra-welfarism) Tsuchiya and Williams argued that,

“(i)t is said that there are two ‘competing views’ on economic evaluation in health care. One is often seen as the ‘theoretically correct’ approach, that is based more firmly within the theory of welfare economics, whilst the other by comparison as some practical but not well formulated collection of rules of thumb (p.22)” (Tsuchiya and Williams, 2001)

The ways in which this collection of rules of thumb (termed extra-welfarist approaches) differs from welfarism and welfare economics more generally were discussed. The focus on health (implicit in the extra-welfarism of Culyer) rather than utility has implications in situations in which potential Pareto improvements occur either in terms of health or utility but not in the other. Extra-welfarism, Communitarianism and Empirical Ethics were outlined as

alternative sets of rules of thumb; only the first of these has been operationalised and the approach to do so for the other two remains unclear.

The quality-adjusted life year (QALY) was then presented as the dominant extra-welfarist numeraire for measuring (and valuing) health changes. The reason for doing so was to illustrate the orthodox approach to valuation of health outcomes, to which the results of my thesis could be applied.

The description and valuation of health

Chapter 2 looked at how health is described and valued for use in economic evaluation. This was important as, while Chapter 1 identified that the value of a health profile was an aggregation of life and quality of length, it did not discuss how quality of life was actually measured. The standard approaches to valuing health states were then described, first covering the method for scoring individual states. Mainstream valuation techniques using general population stated-preference tasks were outlined. For each of the Time Trade-Off, the Standard Gamble and Visual Analogue Scales, some serious theoretical concerns were raised. While the first two of these are based on the concept of trade-offs, other issues (time preference, attitude to risk, respondent burden) impinge on the reliability of the valuations these tools provide for individual health states. These issues are the motivating factor behind the use of the discrete choice experiment in Chapters 5 and 6 of the thesis.

The major health-related quality of life instruments were described and appraised independent of the methods for valuing their constituent health states. The focus was on the tension between descriptive ability and size of the instrument, with each instrument selecting a point of this continuum. While criticism of the relatively small instruments such as the EQ-5D was identified (Hawthorne, et al., 2001), the cost of moving away from it lies in one or both of increasingly correlated dimensions and difficulty in valuing all states within the algorithm. Thus, I concluded that there was no gold standard among the existing suite of instruments, nor will such a gold standard be reached because of this tension. This is unfortunate as it leaves the results from economic evaluations dependent on the instrument used, and thus prone to gaming.

Finally, the techniques required to impute values for other health states not directly valued by respondents were discussed. The need for this imputation increases as the instrument increases in size. Some serious issues with the conventional approaches were identified. Assuming additivity or multiplicativity are both strong assumptions. Both are used by

different teams undertaking this type of research, and the importance of making this kind of constraint on preferences is rarely adequately tested. The chapter concluded by arguing that there might be scope for non-parametric techniques to be used (Kharroubi, et al., 2007; Kharroubi, et al., 2010).

Discrete choice experiments

Chapters 3 and 4 presented the discrete choice experiment as a possible method for exploring some of the questions raised in the previous two chapters. The foundation of discrete choice experiments in Random Utility Theory was introduced (McFadden, 1974; Thurstone, 1927a).

The unusual nature of the trade-off between quality of life and life expectancy required for QALY weights was discussed in the context of Flynn's critique of existing attempts to use ordinal data to explore the issue (Flynn, 2010; McCabe, et al., 2006; Salomon, 2003). While interactions can be considered in a simple additive model, the data had to be constrained to meet the zero-condition in which quality of life is only of value in the context of a non-zero life expectancy, and utility tends to zero as life expectancy tends to zero (Bleichrodt, et al., 1997). Thus, an unusual utility function was specified in which time enters as a main effect and all other characteristics enter only interacted with time, which would be used in subsequent discrete choice experiments.

Flynn (2010) argued that the non-inclusion of time in the ordinal data collection makes it impossible to generate QALY weights with cardinal properties. Including time alongside other characteristics has recently been advocated (Bansback, et al., 2012); however, it is not yet the dominant approach (Hakim and Pathak, 1999; Stolk, et al., 2010), despite being essential for the construction of QALYs.

Chapter 3 then discussed the appropriate tools for investigating response heterogeneity. Using a recent study by Fiebig *et al.* (2010), a series of increasingly relaxed logit models were described for use in subsequent chapters. Thus, a general structure for imposing increasingly relaxed approaches to preferences was outlined, culminating in the generalised multinomial logit model (Fiebig, et al., 2010), which nests each of the preceding models. While Swait and Louviere (1993) argue that accounting for heterogeneity (both scale and preference) is important even if we are only interested in the mean response, the chapter suggested that the mean response may not differ substantially under increasingly sophisticated approaches to modelling heterogeneity, thus limiting their usefulness in generation of population-level

QALY weight algorithms. This issue was left open, to be addressed using the data in chapters 4 and 5.

Chapter 3 then considered the tools for estimating welfare changes based on the results of a DCE. The leading contenders were identified as the marginal rate of substitution and the compensating variation, and the chapter concluded that both had merit under certain circumstances, a result similar to that of Lancsar *et al.* (2007).

Before addressing the empirical components of this thesis, the issue of designing the experiment was then discussed. The principles underpinning regular fractional factorial designs were described, particularly the necessity for the analyst to pre-specify which effects were of greatest interest. Techniques for constructing discrete choice experiments were addressed particularly the approaches of Huber and Zwerina (1996), the L^{MA} , and Street and Burgess (Street, et al., 2005; Street and Burgess, 2007). The chapter addressed the tools for evaluating pre-existing designs, particularly the construction of Λ -, B- and C-matrices, and the choice of maximisation strategies underpinning the leading approaches to experiment efficiency measurement.

DCE 1 – Valuing the SF-6D health states

Chapter 4 used the discrete choice experiment developed in Chapter 3 to provide an alternative valuation approach to states within the SF-6D, described in Chapter 2. The use of design principles maximised the information provided by a fixed sample size. The modelling of heterogeneity allowed the explanation of patterns of responses. The Generalised Multinomial Logit model was preferred under both the Akaike and Bayesian Information Criteria (Akaike, 1974; Schwarz, 1978), although the importance of using more advanced techniques for modelling heterogeneity when the main interest lies in the mean response was questioned.

The results from the DCE reflected the generally monotonic nature of the SF-6D instrument, with one minor exception. Therefore, levels designed to reflect increasingly poorer levels of health were valued increasingly poorly. This was true under a range of specifications of the utility function. QALY weights were derived for the 18,000 SF-6D states, which are useable in economic evaluation.

The chapter considered whether response patterns differed according to observable demographic characteristics of respondents. For example, do men and women value health in

the same way? The results were generally equivocal; while the p-values from the likelihood ratio tests suggested that pooling across gender, chronic conditions or age cannot be accepted, differences in the coefficients in the different groups often did not display clear patterns.

DCE 2 – Equity weights for economic evaluation

Chapter 5 used the DCE approach to evaluate an assumption built into the QALY model described in Chapter 1, namely the assumption that the Social Welfare Function over health is linear (meaning it is inequality-neutral) and symmetrical around the line of equality (meaning it ignores individual characteristics of the person receiving health gain). The evidence to date suggests the QALY model is a very simplified version of true preferences,

“Rather than being linear in quality and length of life, it would seem that social value diminishes in marginal increments of both. And rather than being neutral to the characteristics of people other than their propensity to generate QALYs, the social value of a health improvement seems to be higher if the person has worse lifetime health prospects and higher if that person has dependents. In addition, there is a desire to reduce inequalities in health.” (Dolan, et al., 2005) (p.197)

In some regards, the discussion of what equity consists of matches the discussion in Chapter 1 regarding what the best outcome regarding what the primary outcome of healthcare should be. Thus, it is inevitable that the equity of a situation depends on what dimension an individual believes should be equalised. While this is contentious, a decision had to be made for analysis, and I decided to focus on health outcome as the outcome of interest for the chapter.

A DCE was designed, piloted and run using a representative online panel of Australian residents, and the methods described in Chapter 3 were applied to the data. The respondent was faced with two hypothetical health programs, benefitting different groups by different amounts. Thus, the linearity of utility with respect to health gain could be tested, as could the impact of the characteristics of the hypothetical patients. In addition, a sub-group analysis was undertaken identifying which survey respondent characteristics impacted on which valuations. For example, do smokers systematically differ from non-smokers in terms of how they value health gains accruing to smokers?

The results from the DCE suggested that the average respondent was willing to trade-off some of the total (sum) health of the population to bring about a more equal distribution of

health. In addition, the characteristics of the person receiving the health gain seemed to matter. There was support for the concept of the Fair Innings (Williams, 1997), in that health gains accruing to respondents expected to live 75 years at baseline were valued relatively low. In addition, the average respondent was willing to discriminate in favour of carers and non-smokers. The sub-group analysis suggested that respondents tended to discriminate in favour of (hypothetical) people with similar characteristics to the respondent. This was true when the sub-group analysis was undertaken by smoking status, gender and carer status.

The chapter concluded by discussing whether it is possible and worthwhile to formally incorporate these kinds of preferences into economic evaluation. To do so, I think that three things need to be true:

1. Societal preferences ought to be some aggregation of individual preferences
2. The results from my DCE truly represent preferences, and that the experiment captures all important characteristics
3. That a formal integration of equity concerns into economic evaluation is preferable to the current, more informal approach, in which cost-effectiveness and equity concerns are presented in parallel and the decision maker balances the two

My view is that there is enough uncertainty in each of these points to entail that the formal integration of equity weights into economic evaluation should not be undertaken. Point 2 is clearly a source of uncertainty that could be minimised or removed through further research; it remains moot whether points 1 and 3 reflect a convincing argument for the exclusion of this type of equity weight by themselves. However, I believe that they do not as standard economic evaluation, assuming an equity weight of 1 applying to all individuals, is similarly problematic (albeit more easily ignored).

Summary of importance

The analyses conducted in this thesis provide a number of important results and advances which can be of use in both methodological and applied discussions in the field. The use of the DCE to augment the framework of economic evaluation has been demonstrated in two different settings. A set of weights for the SF-6D has been constructed which can be used in economic evaluation immediately, and have the considerable strength of moving away from the Standard Gamble which, while well-grounded in utility theory, imposes strong assumptions about the structure of individuals' utility functions. A quantitative analysis of public views

with regards to equity was undertaken. This is an important advance over much of the existing literature, as it considers a range of often correlated factors and manages to draw out the impact of each on the value we place on health gain. The consideration of response heterogeneity offers an advance in the field too. It appears likely that the mean response under the various ways of considering heterogeneity remains stable; however, there are important conclusions reached about how people disagree, and what people are thinking when they answer this type of question. The tendency for people to prioritise programs that benefit people like themselves illustrates the need for a sensible approach to sampling the general population when considering contentious topics, and questions whether we should take public responses to stated preference surveys at face value.

Some future directions

One clear conclusion from this thesis is that, while the current approach to measuring and aggregating health outcomes for economic evaluation represents a reasonable approximation in many circumstances, there are other settings in which health and ill-health are not adequately captured (such as the cystic fibrosis example presented in the introductory chapter). Some of these which are outlined below represent future directions for this research.

The first is the occurrence of very short periods of ill-health (or poor quality of life). Under the QALY model, the disutility of a period of time with poor quality of life (e.g. through pain) is proportional to the length of the period. However, this may not be the case. An example might be the use of dental anaesthesia; the period of time in poor health (i.e. extreme pain) is so short that the QALY loss would be negligible, and hence the societal willingness to pay would be very low. However, quite rightly, we as a society do fund these types of interventions. The question is whether we should put the QALY model to one side when considering these types of issues (which themselves are difficult to define) or whether the QALY model should adjust to reflect these issues. This is a question clearly amenable to the DCE approach used in this thesis. Brief periods of ill-health could be nested in longer periods of relatively better health, and the value placed on eliminating this short quality of life shock could be explored.

A similar issue is that of end of life care. Under the QALY model, the capacity to benefit for individuals in a palliative care setting is low. In this case, is this spending rational? Should we be willing to sacrifice average health to provide expensive interventions to individuals in this setting? It is clearly an unpleasant issue, but one in which some would argue that the QALY

model is inadequate. Becker and colleagues (2007) have argued that end-of-life care is undervalued within a conventional QALY-type approach. While, there is likely to be general agreement that decisions should respect patients' choices around end of life care, in economic evaluation, it is assumed that the trade-offs are the same as for any other person (both between aspects of quality of life, and between quality of life and life expectancy); thus the specific circumstances surrounding end of life decision making are not considered. The willingness (or otherwise) of people to undergo difficult and painful interventions to potentially extend their lives by a short (but valuable) period of time is an important topic, that has not been specifically addressed in the literature to date. Again, this is a question to which this type of analysis might be adapted.

The use of DCEs to value health states is becoming more common place (Bansback, et al., 2012; Viney, et al., 2011a). However, the pool of multi-attribute utility instruments that these techniques can potentially be applied to is increasing over time. A good example is the 5-level EQ-5D (Herdman, et al., 2011; Janssen, et al., 2008b). In response to the relative insensitivity of the existing 3-level EQ-5D, the Euroqol group has developed a five level version. While the addition of levels to each dimension is likely to necessitate a larger design for the experiment (if a comparable set of effects are to be estimated), the techniques developed here are clearly relevant to the developing valuation work in this new subfield.

Appendices

Appendix 1: HUI Mark 3	252
Appendix 2: The Assessment of Quality of Life instrument	255
Appendix 3: Final SF-6D DCE Design	258
Appendix 4: SF-6D DCE Screen Shots	262
Appendix 5: RE Probit and RE Logit Results under a Non-Linear Utility Function	272
Appendix 6: SF-6D DCE Subgroup Analysis (Gender)	274
Appendix 7: SF-6D DCE Subgroup Analysis (Age)	276
Appendix 8: SF-6D DCE Subgroup Analysis (Chronic Conditions)	278
Appendix 9: Equity Weights Experiment	280
Appendix 10: Equity Weights for Economic Evaluation DCE	283
Appendix 11: Equity Weights gender subgroup analysis	290
Appendix 12: Equity Weights smoker subgroup analysis	293
Appendix 13: Equity Weights carer status subgroup analysis	295
Appendix 14: Variance Covariance Matrices (Utility Function A)	297
Appendix 15: Variance Covariance Matrices (Utility Function B)	299

Appendix 1: HUI Mark 3

Dimension	Level	
Vision	1	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses.
	2	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses.
	3	Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses.
	4	Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses.
	5	Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses.
	6	Unable to see at all.
Hearing	1	Able to hear what is said in a group conversation with at least three other people, without a hearing aid.
	2	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people.
	3	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people, with a hearing aid.
	4	Able to hear what is said in a conversation with one other person in a quiet room, without a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.
	5	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.
	6	Unable to hear at all.
Speech	1	Able to be understood completely when speaking with strangers or friends.
	2	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well.
	3	Able to be understood partially when speaking with strangers or people who know me well.
	4	Unable to be understood when speaking with strangers but able to be understood partially by people who know me well.

	5	Unable to be understood when speaking to other people (or unable to speak at all).
Ambulation	1	Able to walk around the neighbourhood without difficulty, and without walking equipment.
	2	Able to walk around the neighbourhood with difficulty; but does not require walking equipment or the help of another person.
	3	Able to walk around the neighbourhood with walking equipment, but without the help of another person.
	4	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood.
	5	Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood.
	6	Cannot walk at all.
Dexterity	1	Full use of two hands and ten fingers.
	2	Limitations in the use of hands or fingers, but does not require special tools or help of another person.
	3	Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person).
	4	Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools).
	5	Limitations in use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools).
	6	Limitations in use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools).
Emotion	1	Happy and interested in life.
	2	Somewhat happy.
	3	Somewhat unhappy.
	4	Very unhappy.
	5	So unhappy that life is not worthwhile.
Cognition	1	Able to remember most things, think clearly and solve day to day problems.
	2	Able to remember most things, but have a little difficulty when trying to think and solve day to day problems.
	3	Somewhat forgetful, but able to think clearly and solve day to day problems.
	4	Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems.
	5	Very forgetful, and have great difficulty when trying to think or solve day to day problems.
	6	Unable to remember anything at all, and unable to think or solve day to day problems.

Pain	1	Free of pain and discomfort.
	2	Mild to moderate pain that prevents no activities.
	3	Moderate pain that prevents a few activities.
	4	Moderate to severe pain that prevents some activities.
	5	Severe pain that prevents most activities.

Appendix 2: The Assessment of Quality of Life instrument

Illness	
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
Independent living	
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
Social relationships	
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
Physical senses	
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 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
Psychological well-being	
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

Appendix 3: Final SF-6D DCE Design

PF	RL	SF	PA	MH	VI	DUR	PF	RL	SF	PA	MH	VI	DUR
3	1	3	3	1	3	3	1	2	2	2	4	2	3
5	2	3	0	3	3	3	1	1	3	4	2	2	2
0	0	1	0	1	3	3	5	1	2	1	2	4	3
2	0	3	2	4	4	2	3	0	1	1	3	2	2
2	2	0	2	3	1	1	1	3	0	0	4	2	1
4	3	1	0	4	4	4	1	1	3	4	2	2	2
3	2	1	3	2	4	0	2	3	4	0	4	0	0
2	2	3	5	4	1	0	2	1	4	2	0	2	6
0	0	2	0	2	2	2	5	1	0	0	0	3	1
3	3	2	1	1	3	2	4	0	1	3	0	1	3
1	0	4	1	2	3	0	0	3	2	0	4	1	0
1	3	2	1	0	4	1	1	0	0	1	2	1	5
3	1	3	3	1	3	3	2	0	4	4	0	1	3
5	2	3	0	3	3	3	4	0	0	4	4	0	4
2	1	3	5	4	1	0	3	2	3	4	3	0	2
5	3	3	5	0	3	4	5	3	0	0	2	1	6
2	0	1	1	0	1	1	5	0	1	1	0	0	5
5	2	0	4	2	1	5	4	3	0	3	0	0	5
5	2	0	5	1	0	2	4	3	4	2	2	3	2
3	2	3	4	3	0	2	3	2	0	1	4	4	6
0	2	2	1	3	1	1	3	3	4	4	1	1	2
4	0	2	0	1	0	1	5	3	4	1	2	2	0
5	3	3	3	1	4	1	4	0	2	0	1	0	1
3	0	4	1	2	0	3	1	2	2	2	4	2	3
3	2	4	4	1	4	4	0	2	3	3	0	2	4
1	1	4	5	2	1	3	2	2	0	5	1	4	3
3	1	3	5	0	3	6	3	3	1	3	3	1	6
1	3	2	3	4	3	4	4	1	0	3	2	3	3
0	2	3	3	0	2	4	3	0	0	0	1	0	4
1	0	0	1	2	1	5	0	1	2	4	1	4	5
0	3	1	3	3	3	6	4	3	2	4	1	1	2
1	3	0	2	0	4	6	5	2	1	4	1	2	6
4	2	1	3	1	2	0	3	0	0	2	4	1	0
1	1	4	5	2	1	3	0	0	1	0	1	3	3
2	3	0	1	2	2	6	1	0	2	0	0	0	6
3	0	1	0	0	2	3	0	2	2	3	4	4	3
0	1	3	0	4	1	3	4	1	0	3	2	3	3
1	0	1	2	4	0	4	4	1	4	1	2	1	4
0	3	4	2	1	4	5	0	1	3	0	4	1	3
0	1	0	2	0	0	1	2	2	1	4	1	3	1
0	2	1	1	0	0	3	5	1	1	5	4	4	1
1	3	3	5	3	4	5	4	0	4	3	4	3	5
5	2	4	2	1	4	0	2	2	1	0	0	3	2
0	0	3	4	3	2	1	0	3	4	0	0	2	2
5	2	4	3	3	2	5	4	3	1	5	3	0	3
4	3	3	2	1	1	3	5	1	2	1	2	4	3
5	3	3	5	0	3	4	2	1	4	3	1	1	4
4	1	4	2	3	0	4	2	2	0	5	1	4	3

4	1	3	1	0	3	1	0	2	1	1	0	0	3
2	1	1	1	2	3	2	3	2	4	4	0	1	3
1	3	1	2	3	1	2	4	3	1	0	4	4	4
0	0	4	4	4	3	1	1	2	4	3	0	2	1
4	1	1	1	4	2	5	3	0	1	0	0	2	3
0	1	2	5	2	3	0	1	1	2	2	4	0	5
0	1	2	5	2	3	0	0	0	0	3	3	2	6
1	2	3	0	2	4	6	4	0	2	3	1	4	0
3	1	1	2	4	4	1	4	3	2	4	1	1	2
2	0	3	2	1	3	6	5	2	4	1	4	0	6
1	1	1	3	4	3	2	0	2	2	1	3	1	1
5	0	2	2	3	1	1	5	0	4	0	1	3	4
1	0	4	2	0	4	4	1	1	4	1	3	0	3
3	1	1	2	4	4	1	3	3	0	5	2	3	1
4	2	3	5	4	2	5	1	0	1	2	3	3	5
5	0	4	0	4	1	6	1	2	1	5	1	0	6
3	0	2	2	0	2	3	1	2	4	5	0	4	4
5	3	0	1	4	3	3	5	3	3	2	2	0	3
4	3	3	2	1	1	3	1	3	3	1	1	1	4
4	1	3	2	2	2	6	5	0	0	3	0	0	2
3	3	3	3	2	0	5	5	2	1	2	0	1	5
2	3	2	5	2	2	5	3	1	1	0	3	4	0
2	0	2	4	2	3	6	3	3	1	3	3	1	6
2	1	2	1	3	3	4	3	0	0	1	4	4	2
0	0	3	4	3	2	1	3	2	4	3	2	0	1
2	1	1	1	2	3	2	1	3	2	1	0	4	1
5	2	4	3	3	2	5	2	0	4	4	0	1	3
0	2	0	0	3	2	3	4	1	4	2	3	0	4
5	3	2	4	4	2	1	3	3	0	5	2	3	1
5	1	2	5	1	0	2	1	2	2	5	3	1	4
1	3	3	1	1	1	4	4	1	1	1	4	2	5
1	3	1	4	4	3	0	3	2	2	1	3	2	0
5	3	0	1	4	3	3	0	2	1	5	0	1	4
0	0	3	1	4	1	0	3	1	3	2	1	2	0
0	3	4	1	0	4	2	1	2	3	0	1	0	2
0	3	3	1	1	1	5	4	3	0	3	0	0	5
0	0	1	3	3	4	5	2	0	3	3	0	2	5
5	0	2	2	3	1	1	2	3	2	3	3	4	4
0	0	3	3	4	0	4	4	2	2	4	0	3	4
0	1	2	5	2	2	2	3	3	3	2	0	2	4
4	0	2	3	1	4	0	0	3	4	2	1	3	6
2	1	1	5	0	4	0	4	2	3	4	0	3	0
3	1	4	0	3	3	5	4	2	3	5	4	2	5
2	2	1	0	0	3	2	5	1	4	3	4	2	2
2	3	0	1	4	2	4	0	0	3	3	4	0	4
1	3	2	3	4	3	4	2	0	2	1	1	2	3
4	3	4	5	3	2	0	2	3	4	0	4	0	0
3	0	2	3	1	2	1	5	1	0	0	0	3	1
1	2	3	0	2	4	6	5	1	0	4	3	3	0
3	2	4	4	1	4	4	4	1	4	1	2	1	4
2	2	2	2	4	3	3	1	3	0	3	2	4	3
1	1	3	4	3	0	0	4	2	1	3	1	2	0

1	0	4	1	2	3	0	1	1	3	4	3	0	0
2	0	0	4	1	0	0	3	3	4	5	0	3	5
1	1	1	3	4	3	2	4	2	2	2	2	1	2
2	0	0	4	1	0	0	3	3	2	1	1	3	2
1	1	0	5	1	2	1	2	2	4	4	4	1	5
0	0	4	4	4	3	1	2	2	4	4	4	1	5
2	0	3	2	4	4	2	0	1	1	3	2	1	2
2	3	2	3	3	0	6	3	0	2	4	2	4	5
1	2	3	0	1	0	2	1	3	0	0	4	2	1
1	0	1	2	3	3	5	5	0	1	1	0	0	5
5	1	2	5	1	0	2	1	2	4	3	0	2	1
0	1	0	0	1	4	4	3	2	0	2	3	3	4
0	2	1	5	3	4	5	5	3	3	3	1	4	1
2	2	4	5	1	0	1	4	3	1	4	2	0	1
2	3	2	5	2	2	5	5	2	1	2	0	1	5
5	1	0	4	3	3	0	4	1	2	4	0	4	6
5	1	1	5	4	4	1	3	2	4	3	2	0	1
2	2	0	2	3	1	1	0	3	2	2	0	0	0
3	1	1	0	3	4	0	1	0	3	4	3	4	5
3	3	4	5	0	3	5	1	1	4	4	3	4	3
0	1	0	4	3	0	2	5	3	3	5	0	3	2
0	2	0	0	3	2	3	2	3	3	4	3	0	3
3	3	3	2	0	2	4	4	0	0	2	3	4	2
0	2	1	5	4	3	6	1	2	0	3	0	1	0
5	0	4	0	4	1	6	4	1	3	2	2	2	6
0	3	4	2	1	3	6	2	3	0	1	2	2	6
4	2	3	1	3	4	1	2	3	1	0	2	0	1
3	3	3	5	4	1	1	5	2	2	0	3	2	4
4	1	4	0	3	1	1	1	0	3	3	2	3	1
4	2	3	1	3	4	1	3	3	3	5	4	1	1
0	0	3	4	2	2	4	1	3	0	2	0	4	6
5	0	1	2	2	4	4	0	3	4	2	1	4	5
3	0	2	4	2	4	5	1	2	0	3	0	1	0
0	3	2	2	0	0	0	5	2	4	2	1	4	0
5	2	4	1	4	0	6	2	1	0	5	2	1	4
5	1	2	3	1	1	5	0	3	0	4	0	4	5
0	2	1	5	0	1	4	3	2	1	1	2	0	4
5	0	4	0	1	3	4	1	1	0	5	1	2	1
0	0	3	4	2	2	4	3	1	0	5	4	1	6
4	3	4	5	3	2	0	5	1	3	1	0	4	0
5	0	3	3	3	1	6	3	2	0	1	4	4	6
5	1	2	3	1	1	5	2	0	1	1	0	1	1
0	1	4	1	1	0	6	2	3	2	3	3	0	6
3	1	0	5	4	1	6	4	2	2	0	2	4	6
4	1	4	0	3	1	1	0	1	0	2	0	0	1
3	1	3	5	0	3	6	5	3	4	1	2	2	0
3	1	2	4	0	1	6	5	0	0	3	0	0	2
0	2	4	2	2	0	1	0	3	4	0	0	2	2
2	2	3	0	2	4	5	2	1	1	2	1	2	5
5	3	2	5	3	0	0	1	0	4	0	2	2	0
1	3	1	2	3	1	2	2	1	2	1	3	3	4
1	1	2	2	4	0	5	2	1	0	5	2	1	4

3	0	2	2	0	2	3	3	2	4	4	0	1	3
3	1	2	0	4	0	5	5	2	0	4	2	1	5
0	3	4	1	0	4	2	4	2	0	5	4	2	2
5	1	1	4	2	2	4	4	2	2	4	0	3	4
1	3	1	4	1	2	6	0	1	0	0	1	4	4
5	3	2	4	4	2	1	4	0	4	0	3	4	6
0	0	3	1	4	1	0	1	0	4	0	2	2	0
2	2	3	0	2	4	5	5	3	1	4	4	4	3
4	3	0	0	3	3	5	2	0	2	1	1	2	3
3	2	1	3	2	4	0	4	0	1	0	1	1	0
2	1	1	5	0	4	0	0	2	0	2	2	3	0
5	3	3	5	0	3	2	4	0	0	4	4	0	4
4	0	0	2	3	4	2	4	3	1	4	2	0	1
0	1	1	3	2	1	2	1	2	2	0	0	1	2
0	3	1	3	3	3	6	2	1	4	2	0	2	6
3	2	0	2	3	3	4	2	1	4	3	1	1	4
3	1	3	2	1	2	0	4	2	3	4	0	3	0
4	3	4	2	2	3	2	1	2	2	0	0	1	2
2	1	0	3	0	0	0	1	3	1	4	4	3	0
1	1	4	1	3	0	3	5	3	1	4	4	4	3
1	2	0	1	1	3	5	0	2	4	2	2	0	1
0	1	4	1	1	0	6	2	1	3	0	0	2	5
1	1	4	4	3	4	3	3	2	1	1	2	0	4
4	2	2	0	2	4	6	4	2	3	1	4	0	6
4	2	2	2	2	1	2	2	3	3	0	3	4	2
4	1	0	5	1	2	3	2	2	2	2	4	3	3
2	3	1	0	2	0	1	4	3	0	0	3	3	5
0	1	0	4	3	0	2	2	2	4	3	4	4	2

Appendix 4: SF-6D DCE Screen Shots

Page 1 – Welcome

Welcome

We are inviting you to participate in a study designed to gain an understanding of people's opinions about quality of life, and how people value this and length of life. This will help decision makers in Australia to focus on the areas that Australians value highest. Your responses to hypothetical scenarios will be used to help us to understand what is most important to people in making decisions about different health care treatments which aim to improve quality of life or length of life or both. The study is being undertaken by the Centre for Health Economics Research and Evaluation at the University of Technology, Sydney (UTS).

next

Page 2 - Introduction

The survey contains three sections.

Section A briefly introduces the method we will use to describe health, and ask you to rate your own health.

Section B contains 15 questions. In each question you will be shown three possible health situations. You will then be asked to choose which you would prefer to experience. These profiles do not represent particular conditions and have been made up for the purpose of this exercise. Each option is described in terms of how good your health will be (the different aspects of quality of life, such as mobility, pain, etc) and how long you would expect to live with the condition. In each case you are asked to imagine that the length of time represents the rest of your life (i.e. at the end of the period, you die).

Section C contains questions about you, like your age, which will allow us to apply the results of this study to the population as a whole.

Section D is a brief feedback form, identifying where this survey could be improved.

prev next

Page 3 – Willingness to participate

Your participation in the study is completely voluntary. You are not obliged to participate and may stop the study at any time. Your responses to this survey are strictly confidential and at no time will the answers you give be linked to your identity. To complete this survey, we would expect someone to take approximately 15 minutes.

If you would like to speak to someone about the study, or the survey itself, please call Richard Norman at the Centre for Health Economics Research and Evaluation on 02 9514 4732.

Are you willing to participate?

Yes

No

next

Page 4 – Describing own health (SF6D)

Section A: Describing Your Health

In this survey, we will describe health in a particular way, describing a life using a number of areas such as pain, mental health and vitality. To familiarise yourself with this approach, please answer these questions.

For each of these six areas, which of these best describes your current health?

Which of these best describes your current situation?

- Your health does not limit you in vigorous activities
- Your health limits you a little in vigorous activities
- Your health limits you a little in moderate activities
- Your health limits you a lot in moderate activities
- Your health limits you a little in bathing and dressing
- Your health limits you a lot in bathing and dressing

Which of these best describes your current situation?

- You have no problem with your work or other regular daily activities as a result of your physical health or any emotional problems
- You are limited in the kind of work or other activities as a result of your physical health
- You accomplish less than you would like as a result of emotional problems
- You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems

Which of these best describes your current situation?

- Your health limits your social activities none of the time
- Your health limits your social activities a little of the time
- Your health limits your social activities some of the time
- Your health limits your social activities most of the time
- Your health limits your social activities all of the time

Which of these best describes your current situation?

- You have no pain
- You have pain but it does not interfere with your normal work (both outside the home and housework)
- You have pain that interferes with your normal work (both outside the home and housework) a little bit
- You have pain that interferes with your normal work (both outside the home and housework) moderately
- You have pain that interferes with your normal work (both outside the home and housework) quite a bit
- You have pain that interferes with your normal work (both outside the home and housework) extremely

Which of these best describes your current situation?

- You feel tense or downhearted and low none of the time
- You feel tense or downhearted and low a little of the time
- You feel tense or downhearted and low some of the time
- You feel tense or downhearted and low most of the time
- You feel tense or downhearted and low all of the time

Which of these best describes your current situation?

- You always have a lot of energy
- You usually have a lot of energy
- You sometimes have a lot of energy
- You rarely have a lot of energy
- You never have a lot of energy

[prev](#) [next](#)

Page 5 - Describing own health (EQ-5D)

Thank you. Before we begin asking you questions, we would like you to select which of the following options in each question best reflects your current health

Which of these best describes your current situation?

- You have no problems in walking about
- You have some problems in walking about
- You are confined to bed

Which of these best describes your current situation?

- You have no problems with self-care
- You have some problems washing and dressing myself
- You are unable to wash and dress myself

Which of these best describes your current situation?

- You have no problems with performing your usual activities
- You have some problems with performing your usual activities
- You are unable to perform your usual activities

Which of these best describes your current situation?

- You have no pain or discomfort
- You have moderate pain or discomfort
- You have extreme pain or discomfort

Which of these best describes your current situation?

- You are not anxious or depressed
- You are moderately anxious or depressed
- You are extremely anxious or depressed

[prev](#) [next](#)

Page 6 - Example

Section B: Making Choices Between Options

You will now be presented with 15 choices. In each situation, you will see three scenarios.

The first two describe two different combinations of health and life expectancy: that is how long you would live under this scenario and what your health related quality of life would be. The third scenario in each choice is death – which is presented because for some health states the quality of life is such that some people may prefer not to live at all with this quality of life.

Consider the example below. The blue box describes your health under Scenario A. Your life expectancy under this health state is shown as one year (meaning you will live for one year and then die). This Option is to be compared with Scenario B and death. For Scenario B, your health is described in the green box, and your life expectancy is 10 years (meaning you will live for 10 years and then die).

Your task is to decide which of the three scenarios is better. In each choice we will ask you to tell us first which of the three scenarios is the best – that is, the one you would most prefer to experience out of the three. We will then ask you to tell us which of the remaining two scenarios is the worst.

If you had to choose between the following scenarios:

	State 1	State 2	Immediate death
Physical Functioning	Your health limits you a lot in moderate activities	Your health limits you a lot in bathing and dressing	
Role Limitation	You are limited in the kind of work or other activities as a result of your physical health	You accomplish less than you would like as a result of emotional problems	
Social Functioning	Your health limits your social activities some of the time	Your health limits your social activities none of the time	
Pain	You have no pain	You have pain that interferes with your normal work (both outside the home and housework) quite a bit	
Mental Health	You feel tense or downhearted and low all of the time	You feel tense or downhearted and low some of the time	
Vitality	You always have a lot of energy	You usually have a lot of energy	
Duration	15 years, followed by death	15 years, followed by death	
Which option is the best?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Which option is the worst?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[prev](#) [next](#)

There are no right or wrong answers. Some of the health state and life expectancy combinations may be difficult for you to imagine - just do the best you can. We are interested in your views, because it will help us to understand what aspects of quality of life are most important to people.

Please make sure you consider all aspects of the health state and the life expectancy for each scenario. If you need the instructions again whilst working through the questions click on the




Let's start the questions now.

[prev](#) [next](#)

Pages 7-22 – The Choice Experiment

If you had to choose between the following states:

			
	State 1	State 2	Immediate death
Physical Functioning	Your health limits you a little in moderate activities	Your health limits you a lot in bathing and dressing	
Role Limitation	You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems	You have no problem with your work or other regular daily activities as a result of your physical health or any emotional problems	
Social Functioning	Your health limits your social activities some of the time	Your health limits your social activities some of the time	
Pain	You have pain that interferes with your normal work (both outside the home and housework) moderately	You have pain that interferes with your normal work (both outside the home and housework) a little bit	
Mental Health	You feel tense or downhearted and low most of the time	You feel tense or downhearted and low most of the time	
Vitality	You never have a lot of energy	You usually have a lot of energy	
Duration	12 years, followed by death	2 years, followed by death	
Which option is the best?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Which option is the worst?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[prev](#) [next](#)

Page 23 – Thanks and methods

You have now finished the 15 questions. Thank you.

In general, which of this statements best represents your approach to answering the task?

- I considered all of the areas of health and life expectancy before making my decision.
- I focused on some of the areas and made my decision based on these.
- I selected at random.

[prev](#) [next](#)

Page 24 – Basic Information (I)

Section C: Some more information about you

Are you:

Male

Female

How old are you (in years)?

Are you

Aboriginal

Torres Strait Islander

Neither

Both

[prev](#) [next](#)

Page 25 - Basic Information (II)

Country of Birth

Australia

England

New Zealand

Italy

Vietnam

Scotland

Greece

Germany

Philippines

Netherlands

Other (please specify)

[prev](#) [next](#)

Page 26 – Basic Information (III)

What year did you arrive in Australia (if born outside Australia)?

[prev](#) [next](#)

Page 27 – Basic Information (IV)

What is the highest level of education you completed?

- Completed primary
- Completed secondary
- Trade certificate/Diploma
- Bachelor's degree
- Higher degree

Are you currently studying?

- Yes
- No

Please indicate which of these categories best matches your before tax household income.

- Under \$20,000
- \$20,001-\$30,000
- \$30,001-\$40,000
- \$40,001-\$50,000
- \$50,001-\$60,000
- \$60,001-\$70,000
- \$70,001-\$80,000
- \$80,001-\$100,000
- Over \$100,000
- Not stated

[prev](#) [next](#)

Page 28 – Basic Information (V)

What is your current marital status?

Single

Separated/Divorced

Widowed

Married/De facto

Please enter the number of children for each of the age ranges below

	number of children
less than 5 years	<input type="text"/>
aged between 5 and 16 years	<input type="text"/>
older than 16 years?	<input type="text"/>

Page 29 – Basic Information (VI)

In general would you say that your health is excellent, very good, good, fair or poor?

Excellent

Very good

Good

Fair

Poor

Do you have a chronic health condition?

Yes

No

Have you had a health condition requiring hospitalisation in the last 5 years?

Yes

No

Page 30 – Basic Information (VII)

Do you usually work in a paid job?

Yes

No

How many people in your household work in a paid job?

None

One

Two

More than two

[prev](#) [next](#)

Page 31 – Improving the survey and other feedback

Section D: Improving the Survey

Thankyou for completing the survey. We would like to ask you to answer the following question, identifying how easy you found the survey to answer. If you wish to add any details of this, please add this in the box below. Thankyou in advance for any feedback you may have.

In general, how clear was the task that you were asked to complete?

Very difficult

Quite difficult

Neither easy nor difficult

Quite easy

Very easy

If you have any general comments about the survey, please give them here

[prev](#) [next](#)

Appendix 5: RE Probit and RE Logit Results under a Non-Linear Utility Function

	RE Probit		RE Logit	
	Coefficient (Standard error)	p-value	Coefficient (Standard error)	p-value
duration	0.4674 (0.0189)	0.000	0.8153 (0.0335)	0.000
dur_pf2	-0.0250 (0.0114)	0.028	-0.0449 (0.0194)	0.020
dur_pf3	-0.0214 (0.0101)	0.034	-0.0458 (0.0173)	0.008
dur_pf4	-0.0784 (0.0112)	0.000	-0.1407 (0.0194)	0.000
dur_pf5	-0.0634 (0.0115)	0.000	-0.1170 (0.0198)	0.000
dur_pf6	-0.1332 (0.0121)	0.000	-0.2399 (0.0210)	0.000
dur_rl2	-0.0412 (0.0102)	0.000	-0.0665 (0.0173)	0.000
dur_rl3	-0.0102 (0.0093)	0.275	-0.0071 (0.0160)	0.659
dur_rl4	-0.0250 (0.0095)	0.008	-0.0447 (0.0161)	0.006
dur_sf2	-0.0684 (0.0116)	0.000	-0.1209 (0.0201)	0.000
dur_sf3	-0.0350 (0.0106)	0.001	-0.0715 (0.0179)	0.000
dur_sf4	-0.0925 (0.0108)	0.000	-0.1579 (0.0183)	0.000
dur_sf5	-0.0849 (0.0106)	0.000	-0.1475 (0.0181)	0.000
dur_pa2	-0.0747 (0.0121)	0.000	-0.1097 (0.0207)	0.000
dur_pa3	-0.1023 (0.0107)	0.000	-0.1617 (0.0181)	0.000
dur_pa4	-0.0686 (0.0116)	0.000	-0.1169 (0.0197)	0.000
dur_pa5	-0.1349 (0.0109)	0.000	-0.2359 (0.0188)	0.000
dur_pa6	-0.1920 (0.0132)	0.000	-0.3206 (0.0226)	0.000
dur_mh2	-0.0338 (0.0101)	0.001	-0.0625 (0.0173)	0.000
dur_mh3	-0.0237 (0.0103)	0.021	-0.0449 (0.0174)	0.010
dur_mh4	-0.1071 (0.0113)	0.000	-0.1900 (0.0193)	0.000
dur_mh5	-0.1412 (0.0099)	0.000	-0.2479 (0.0170)	0.000
dur_vi2	-0.0501 (0.0120)	0.000	-0.0911 (0.0204)	0.000
dur_vi3	-0.0741 (0.0104)	0.000	-0.1290 (0.0176)	0.000
dur_vi4	-0.0987 (0.0108)	0.000	-0.1701 (0.0184)	0.000
dur_vi5	-0.1659 (0.0106)	0.000	-0.2882 (0.0182)	0.000
constant	-0.0057 (0.0110)	0.607	-0.0086 (0.0184)	0.640
Duration ²	-0.0197 (0.0012)	0.000	-0.0346 (0.0021)	0.000
dur ² _pf2	0.0021 (0.0007)	0.005	0.0038 (0.0013)	0.002
dur ² _pf3	0.0008 (0.0006)	0.217	0.0020 (0.0011)	0.068
dur ² _pf4	0.0031 (0.0007)	0.000	0.0057 (0.0012)	0.000
dur ² _pf5	0.003 (0.0008)	0.000	0.0058 (0.0013)	0.000
dur ² _pf6	0.0061 (0.0008)	0.000	0.0111 (0.0014)	0.000
dur ² _rl2	0.0020 (0.0007)	0.003	0.0034 (0.0011)	0.003
dur ² _rl3	-0.0003 (0.0006)	0.619	-0.0011 (0.0010)	0.288
dur ² _rl4	0.0010 (0.0006)	0.105	0.0017 (0.0010)	0.097
dur ² _sf2	0.0043 (0.0007)	0.000	0.0075 (0.0013)	0.000
dur ² _sf3	0.0024 (0.0007)	0.001	0.0045 (0.0012)	0.000
dur ² _sf4	0.0047 (0.0007)	0.000	0.0080 (0.0012)	0.000
dur ² _sf5	0.0042 (0.0007)	0.000	0.0074 (0.0012)	0.000
dur ² _pa2	0.0042 (0.0008)	0.000	0.0063 (0.0013)	0.000
dur ² _pa3	0.0052 (0.0007)	0.000	0.0081 (0.0012)	0.000
dur ² _pa4	0.0025 (0.0008)	0.001	0.0044 (0.0013)	0.001

dur ² _pa5	0.0058 (0.0007)	0.000	0.0102 (0.0012)	0.000
dur ² _pa6	0.0094 (0.0008)	0.000	0.0156 (0.0014)	0.000
dur ² _mh2	0.0016 (0.0006)	0.013	0.0031 (0.0011)	0.005
dur ² _mh3	0.0010 (0.0006)	0.113	0.0022 (0.0011)	0.045
dur ² _mh4	0.0049 (0.0007)	0.000	0.0088 (0.0012)	0.000
dur ² _mh5	0.0063 (0.0006)	0.000	0.0113 (0.0011)	0.000
dur ² _vi2	0.0031 (0.0008)	0.000	0.0055 (0.0013)	0.000
dur ² _vi3	0.0042 (0.0007)	0.000	0.0072 (0.0012)	0.000
dur ² _vi4	0.0042 (0.0007)	0.000	0.0073 (0.0012)	0.000
dur ² _vi5	0.0084 (0.0007)	0.000	0.0147 (0.0011)	0.000
/lnsig2u	-12.4832 (8.4301)		-12.2185 (8.8422)	
sigma_u	0.0019 (0.0082)		0.0022 (0.0098)	
ρ	0.0000 (0.0000)		0.0000 (0.0000)	
Log likelihood	-8667		-8639	
AIC	17385		17387	
BIC	17789		17799	

* Coefficients in **bold** are statistically significant at the 1% level

Appendix 6: SF-6D DCE Subgroup Analysis (Gender)

	Coefficient	Standard Error	z	P>z	[95% Conf.	Interval]
duration	0.193372	0.006103	31.69	0	0.181411	0.205334
dur_pf2	-0.01062	0.003388	-3.13	0.002	-0.01726	-0.00398
dur_pf3	-0.01738	0.003355	-5.18	0	-0.02396	-0.01081
dur_pf4	-0.0305	0.003279	-9.3	0	-0.03692	-0.02407
dur_pf5	-0.02933	0.003433	-8.55	0	-0.03606	-0.02261
dur_pf6	-0.05858	0.00343	-17.08	0	-0.0653	-0.05185
dur_rl2	-0.01865	0.002981	-6.26	0	-0.0245	-0.01281
dur_rl3	-0.01463	0.002755	-5.31	0	-0.02003	-0.00923
dur_rl4	-0.02199	0.002888	-7.61	0	-0.02765	-0.01633
dur_sf2	-0.00405	0.003246	-1.25	0.212	-0.01041	0.00231
dur_sf3	-0.00348	0.003042	-1.14	0.253	-0.00944	0.002482
dur_sf4	-0.02162	0.002948	-7.33	0	-0.02739	-0.01584
dur_sf5	-0.02168	0.003139	-6.91	0	-0.02783	-0.01553
dur_pa2	-0.01476	0.003446	-4.28	0	-0.02152	-0.00801
dur_pa3	-0.03508	0.003208	-10.93	0	-0.04137	-0.02879
dur_pa4	-0.04066	0.003433	-11.84	0	-0.04739	-0.03393
dur_pa5	-0.05496	0.003292	-16.7	0	-0.06141	-0.04851
dur_pa6	-0.05292	0.003564	-14.85	0	-0.05991	-0.04593
dur_mh2	-0.01462	0.002954	-4.95	0	-0.02041	-0.00883
dur_mh3	-0.01377	0.002978	-4.62	0	-0.0196	-0.00793
dur_mh4	-0.03813	0.003037	-12.56	0	-0.04409	-0.03218
dur_mh5	-0.04981	0.002915	-17.09	0	-0.05553	-0.0441
dur_vi2	0.000966	0.002959	0.33	0.744	-0.00483	0.006766
dur_vi3	-0.01028	0.003192	-3.22	0.001	-0.01654	-0.00403
dur_vi4	-0.03784	0.003086	-12.26	0	-0.04389	-0.03179
dur_vi5	-0.04713	0.003258	-14.46	0	-0.05351	-0.04074
duration_f~e	0.008393	0.009507	0.88	0.377	-0.01024	0.027026
dur_pf2_fem	-0.00669	0.006218	-1.08	0.282	-0.01888	0.005495
dur_pf3_fem	-0.0089	0.006399	-1.39	0.164	-0.02144	0.003638
dur_pf4_fem	-0.01307	0.006047	-2.16	0.031	-0.02492	-0.00122
dur_pf5_fem	-0.00855	0.00612	-1.4	0.162	-0.02054	0.003445
dur_pf6_fem	-0.00921	0.006234	-1.48	0.14	-0.02143	0.003009
dur_rl2_fem	-0.00343	0.005318	-0.64	0.519	-0.01385	0.006997
dur_rl3_fem	-0.00501	0.00507	-0.99	0.323	-0.01495	0.004927
dur_rl4_fem	0.000422	0.005136	0.08	0.935	-0.00964	0.010487
dur_sf2_fem	0.000998	0.00532	0.19	0.851	-0.00943	0.011426
dur_sf3_fem	0.010134	0.005734	1.77	0.077	-0.0011	0.021372
dur_sf4_fem	-0.00026	0.005596	-0.05	0.963	-0.01123	0.010706
dur_sf5_fem	0.008569	0.005349	1.6	0.109	-0.00192	0.019053
dur_pa2_fem	0.00267	0.006136	0.44	0.663	-0.00936	0.014697
dur_pa3_fem	-0.00506	0.005959	-0.85	0.396	-0.01674	0.006621
dur_pa4_fem	-0.00426	0.006133	-0.69	0.487	-0.01628	0.007761
dur_pa5_fem	0.0006	0.006184	0.1	0.923	-0.01152	0.012719
dur_pa6_fem	-0.00225	0.006632	-0.34	0.735	-0.01525	0.010749
dur_mh2_fem	-0.01096	0.005803	-1.89	0.059	-0.02233	0.000417

dur_mh3_fem	0.002236	0.005699	0.39	0.695	-0.00893	0.013406
dur_mh4_fem	-0.00747	0.005576	-1.34	0.18	-0.0184	0.00346
dur_mh5_fem	0.008475	0.005336	1.59	0.112	-0.00198	0.018934
dur_vi2_fem	0.009103	0.005897	1.54	0.123	-0.00246	0.020661
dur_vi3_fem	-0.00546	0.005788	-0.94	0.345	-0.0168	0.005883
dur_vi4_fem	0.008548	0.005694	1.5	0.133	-0.00261	0.019708
dur_vi5_fem	0.001624	0.005618	0.29	0.773	-0.00939	0.012636
constant	0.001809	0.014226	0.13	0.899	-0.02607	0.029692

Appendix 7: SF-6D DCE Subgroup Analysis (Age)

choice	Coefficient	Standard Error	z	P>z	[95% Conf.	Interval]
duration	0.194223	0.005949	32.65	0	0.182562	0.205884
dur_pf2	-0.0132	0.003296	-4.01	0	-0.01966	-0.00674
dur_pf3	-0.01986	0.003208	-6.19	0	-0.02615	-0.01357
dur_pf4	-0.02876	0.003166	-9.08	0	-0.03497	-0.02256
dur_pf5	-0.03274	0.003332	-9.83	0	-0.03927	-0.02621
dur_pf6	-0.0596	0.003318	-17.96	0	-0.06611	-0.0531
dur_rl2	-0.01976	0.002883	-6.85	0	-0.02541	-0.01411
dur_rl3	-0.01412	0.00267	-5.29	0	-0.01935	-0.00889
dur_rl4	-0.02306	0.002809	-8.21	0	-0.02857	-0.01756
dur_sf2	-0.0047	0.003169	-1.48	0.138	-0.01092	0.001506
dur_sf3	-0.00311	0.002966	-1.05	0.294	-0.00893	0.002703
dur_sf4	-0.01991	0.002829	-7.04	0	-0.02545	-0.01436
dur_sf5	-0.02319	0.003051	-7.6	0	-0.02917	-0.01721
dur_pa2	-0.01813	0.003335	-5.44	0	-0.02467	-0.01159
dur_pa3	-0.03461	0.00309	-11.2	0	-0.04066	-0.02855
dur_pa4	-0.04184	0.003383	-12.37	0	-0.04847	-0.03521
dur_pa5	-0.05899	0.003207	-18.39	0	-0.06528	-0.05271
dur_pa6	-0.05402	0.00347	-15.57	0	-0.06082	-0.04722
dur_mh2	-0.00938	0.002826	-3.32	0.001	-0.01492	-0.00384
dur_mh3	-0.01238	0.002846	-4.35	0	-0.01795	-0.0068
dur_mh4	-0.03463	0.002921	-11.85	0	-0.04035	-0.0289
dur_mh5	-0.04817	0.002806	-17.17	0	-0.05367	-0.04268
dur_vi2	-0.00237	0.002851	-0.83	0.405	-0.00796	0.003215
dur_vi3	-0.00878	0.003104	-2.83	0.005	-0.01487	-0.0027
dur_vi4	-0.04117	0.003008	-13.69	0	-0.04706	-0.03527
dur_vi5	-0.04883	0.00317	-15.41	0	-0.05504	-0.04262
duration_old	0.013855	0.010005	1.38	0.166	-0.00576	0.033465
dur_pf2_old	-0.01726	0.00646	-2.67	0.008	-0.02992	-0.0046
dur_pf3_old	-0.02091	0.006701	-3.12	0.002	-0.03405	-0.00778
dur_pf4_old	-0.00883	0.00634	-1.39	0.164	-0.02126	0.003597
dur_pf5_old	-0.02273	0.006352	-3.58	0	-0.03518	-0.01028
dur_pf6_old	-0.01264	0.006507	-1.94	0.052	-0.0254	0.000111
dur_rl2_old	-0.0089	0.005562	-1.6	0.11	-0.0198	0.002004
dur_rl3_old	-0.0041	0.005324	-0.77	0.441	-0.01454	0.00633
dur_rl4_old	-0.00203	0.005411	-0.38	0.707	-0.01264	0.008571
dur_sf2_old	-0.00125	0.005682	-0.22	0.826	-0.01239	0.009886
dur_sf3_old	0.010691	0.006035	1.77	0.076	-0.00114	0.022519
dur_sf4_old	0.004378	0.005826	0.75	0.452	-0.00704	0.015796
dur_sf5_old	0.001904	0.005604	0.34	0.734	-0.00908	0.012888
dur_pa2_old	-0.01154	0.006409	-1.8	0.072	-0.0241	0.001023
dur_pa3_old	-0.00657	0.00626	-1.05	0.294	-0.01884	0.005699
dur_pa4_old	-0.01219	0.006313	-1.93	0.053	-0.02457	0.000179

dur_pa5_old	-0.01527	0.006427	-2.38	0.018	-0.02787	-0.00267
dur_pa6_old	-0.00608	0.006905	-0.88	0.378	-0.01962	0.00745
dur_mh2_old	0.009577	0.006041	1.59	0.113	-0.00226	0.021418
dur_mh3_old	0.009355	0.005988	1.56	0.118	-0.00238	0.021091
dur_mh4_old	0.005624	0.005849	0.96	0.336	-0.00584	0.017088
dur_mh5_old	0.016298	0.005641	2.89	0.004	0.005242	0.027353
dur_vi2_old	-0.00191	0.006126	-0.31	0.756	-0.01391	0.010102
dur_vi3_old	0.000624	0.006001	0.1	0.917	-0.01114	0.012385
dur_vi4_old	-0.00161	0.005894	-0.27	0.785	-0.01316	0.00994
dur_vi5_old	-0.00494	0.005877	-0.84	0.4	-0.01646	0.006574
Constant	0.02681	0.013477	1.99	0.047	0.000395	0.053225

Appendix 8: SF-6D DCE Subgroup Analysis (Chronic Conditions)

	Coefficient	Std. Err.	z	P>z	[95% Conf.	Interval]
duration	0.195847	0.005838	33.54	0	0.184404	0.20729
dur_pf2	-0.01217	0.003181	-3.82	0	-0.0184	-0.00593
dur_pf3	-0.01769	0.003127	-5.66	0	-0.02382	-0.01156
dur_pf4	-0.03089	0.003085	-10.01	0	-0.03693	-0.02484
dur_pf5	-0.03221	0.003216	-10.02	0	-0.03851	-0.02591
dur_pf6	-0.06097	0.003231	-18.87	0	-0.0673	-0.05464
dur_rl2	-0.01944	0.002816	-6.9	0	-0.02496	-0.01392
dur_rl3	-0.01403	0.002592	-5.41	0	-0.01911	-0.00895
dur_rl4	-0.02441	0.002712	-9	0	-0.02973	-0.0191
dur_sf2	-0.00578	0.003061	-1.89	0.059	-0.01178	0.000219
dur_sf3	-0.00463	0.002832	-1.64	0.102	-0.01018	0.000918
dur_sf4	-0.02136	0.002729	-7.83	0	-0.02671	-0.01601
dur_sf5	-0.02461	0.002964	-8.3	0	-0.03042	-0.0188
dur_pa2	-0.0141	0.003226	-4.37	0	-0.02043	-0.00778
dur_pa3	-0.03318	0.002993	-11.08	0	-0.03904	-0.02731
dur_pa4	-0.04047	0.003222	-12.56	0	-0.04679	-0.03416
dur_pa5	-0.0574	0.003095	-18.54	0	-0.06347	-0.05133
dur_pa6	-0.05265	0.003343	-15.75	0	-0.0592	-0.0461
dur_mh2	-0.01293	0.002764	-4.68	0	-0.01835	-0.00751
dur_mh3	-0.01366	0.002754	-4.96	0	-0.01906	-0.00827
dur_mh4	-0.03527	0.002876	-12.27	0	-0.04091	-0.02964
dur_mh5	-0.05145	0.002757	-18.66	0	-0.05686	-0.04605
dur_vi2	-0.00107	0.002721	-0.39	0.695	-0.0064	0.004266
dur_vi3	-0.0091	0.003001	-3.03	0.002	-0.01498	-0.00321
dur_vi4	-0.04262	0.002914	-14.63	0	-0.04833	-0.03691
dur_vi5	-0.0487	0.003093	-15.74	0	-0.05477	-0.04264
duration_c~c	0.028925	0.010834	2.67	0.008	0.007692	0.050158
dur_pf2_ch~c	-0.01868	0.007145	-2.61	0.009	-0.03268	-0.00468
dur_pf3_ch~c	-0.01384	0.00719	-1.92	0.054	-0.02793	0.000257
dur_pf4_ch~c	-0.0205	0.006766	-3.03	0.002	-0.03376	-0.00723
dur_pf5_ch~c	-0.02794	0.006993	-4	0	-0.04164	-0.01423
dur_pf6_ch~c	-0.02337	0.00713	-3.28	0.001	-0.03735	-0.0094
dur_rl2_ch~c	-0.00953	0.006069	-1.57	0.116	-0.02142	0.002368
dur_rl3_ch~c	-0.00425	0.0057	-0.75	0.456	-0.01542	0.006921
dur_rl4_ch~c	-0.0104	0.005807	-1.79	0.073	-0.02178	0.00098
dur_sf2_ch~c	-0.00623	0.00608	-1.02	0.305	-0.01815	0.005685
dur_sf3_ch~c	0.009465	0.006372	1.49	0.137	-0.00302	0.021954
dur_sf4_ch~c	0.000741	0.006274	0.12	0.906	-0.01156	0.013038
dur_sf5_ch~c	-0.00304	0.006116	-0.5	0.619	-0.01502	0.008949
dur_pa2_ch~c	0.003688	0.006993	0.53	0.598	-0.01002	0.017393
dur_pa3_ch~c	-0.0017	0.006947	-0.25	0.806	-0.01532	0.011912
dur_pa4_ch~c	-0.00844	0.007018	-1.2	0.229	-0.02219	0.005321
dur_pa5_ch~c	-0.01179	0.006938	-1.7	0.089	-0.02539	0.001808

dur_pa6_ch~c	-0.00459	0.007556	-0.61	0.543	-0.0194	0.010218
dur_mh2_ch~c	-0.00909	0.006476	-1.4	0.16	-0.02179	0.003602
dur_mh3_ch~c	0.002356	0.006475	0.36	0.716	-0.01034	0.015047
dur_mh4_ch~c	0.001556	0.006291	0.25	0.805	-0.01077	0.013887
dur_mh5_ch~c	0.00454	0.006112	0.74	0.458	-0.00744	0.016519
dur_vi2_ch~c	0.001425	0.006566	0.22	0.828	-0.01144	0.014294
dur_vi3_ch~c	-0.00294	0.006628	-0.44	0.658	-0.01593	0.010054
dur_vi4_ch~c	-0.01287	0.006457	-1.99	0.046	-0.02552	-0.00022
dur_vi5_ch~c	-0.00732	0.006448	-1.13	0.257	-0.01995	0.005322
Constant	0.025302	0.01293	1.96	0.05	-4.1E-05	0.050644

Appendix 9: Equity Weights Experiment

The first seven figures in each 14-digit row refer to the characteristics of the first group of potential respondents, the second seven figures refer to those of the second group of potential respondents.


10001010100011,	00011031101013,	01110021011132,	00100201000123,
00011001011001,	00001001010003,	00110201100020,	11011230010123,
01011201111023,	01101001010030,	10101130000012,	00110321001133,
11110320011102,	11100110010101,	00011211110121,	11110300101131,
10101300110020,	00110311001132,	11100130100112,	01000321000122,
11110020011112,	00001111100001,	10110200001123,	10100100000111,
00010131101103,	00011021110102,	11100130010103,	10001310100001,
11111200011010,	11101110001111,	01101211010011,	01101331010023,
01111231101020,	11010110001101,	11010310001121,	00000221010123,
01001021110003,	01001131000023,	10001010111101,	11111000011030,
10101100000013,	00010101101100,	00101101110020,	00111001001003,
01100221100123,	10001030010000,	00100001110130,	01011031001033,
10011210101011,	01100211001021,	01011131111012,	00110001111110,
00000301100100,	10001020111102,	00111101001013,	00100211000120,
00110131001110,	11000320000102,	10100330000130,	10111210001022,
01010111010011,	10101200110010,	11000120110113,	01001111011111,
00001021010001,	10101230110013,	10000330010132,	11100210100120,
10001210111121,	01001321000002,	01000021000132,	00101301101130,
11011320111033,	01100001010110,	01111221011032,	00110111001112,
01010121010012,	00000301111030,	00011131110113,	10101010000000,
10110020111132,	10001300100000,	11000110000121,	00100211110111,
01001031011103,	01010201001130,	11010110111110,	01010011001111,
01100101100111,	11110310011101,	00010011011100,	01010301010030,
10010000011101,	00000221111022,	11011100010110,	00000311111031,
01001031110000,	01000021110101,	00010001011103,	10101320000031,
01100231010133,	00001201100010,	01111331101030,	11001030000033,
10111200001021,	11010010001131,	01111111011021,	01011321111031,
10110300001133,	00100011110131,	00110201111130,	11101120001112,
11001310000021,	00011311011032,	10001020010003,	01111011101002,
00110001100000,	11111130011003,	10001130111113,	01101311100030,
01100031010113,	11101020100003,	00001311100021,	00000201100130,
11011130001023,	10111300111000,	00000321111032,	01011001111003,
00010131011112,	00100011000100,	00010201011123,	11000300000100,
11111010101000,	11101220001122,	01010311001101,	01111331000133,
00110031100003,	00011301011031,	11101130010023,	10001320010033,
01000131000103,	00101311110001,	11000210011021,	00100031110133,
00011001110100,	11000130110110,	10100330110103,	11110130101110,
00111321111022,	10101220110012,	11000030110100,	01000311110130,
10110010111131,	00101131101113,	01101101010000,	11000310011031,
01111231011033,	10111310111001,	10110300111120,	11010030001133,
10011030101033,	10111020100102,	01100331100130,	11100320010122,
10011130101003,	11011010111002,	11111020101001,	10010220110022,
10100030101003,	00000131111013,	00110221111132,	01110301101133,
10111220001023,	11110120000012,	01010211001131,	10001230111123,
11110010101102,	01111001000100,	10110310111121,	00010111011110,
01001131011113,	00100321110122,	10001000100010,	00111331111023,
01000111110110,	00011211011022,	01100031100100,	01010301111131,
11011300001000,	00100111000110,	01001101000020,	11101130100010,
11101200001120,	11010210111120,	11100330100132,	10111220111032,
11111230011013,	00101021110012,	01001311000001,	01010321001102,
11011010010101,	01010101010010,	01010201010020,	10110230111113,
01100111100112,	01110111101110,	00000011010102,	10110000111130,

01111011011011,	11101330100030,	01110011101100,	11001210110020,
01011011111000,	11000310000101,	01011111111010,	01110101011100,
01100021100103,	00101331000030,	10011220011021,	11011200111021,
00111221001021,	01110031101102,	01011231001013,	11110030011113,
00100021000101,	00011231110123,	00101021000003,	01111311101032,
11000300011030,	11110200011130,	00101201101120,	01001031000013,
11101200100021,	01100321010102,	11101110100012,	10000030100133,
01000031000133,	00110311111101,	11000020110103,	11101120010022,
11100220100121,	11000220000132,	01010211111122,	11101320010002,
10001320111132,	01010311111132,	11110130011123,	11010330111132,
00000031100113,	11111130101012,	11101000100001,	11000120011012,
10010100011111,	00101321101132,	01111121000112,	10111210111031,
01110301011120,	00110011111111,	01111311000131,	00001321010031,
10001300111130,	10010310101101,	10010100110010,	00011320111033,
01100221010132,	01010221111123,	11100320100131,	10110120111102,
00101111000012,	00011201011021,	01011321001022,	01001001110001,
01100211100122,	10100230101023,	11001020000032,	00000231100133,
00010021011101,	10101110110001,	11011310010131,	11000200000130,
10000300100120,	10110100001113,	10000020010101,	10110120001111,
01000131110112,	00000031010100,	00000021010103,	01111101000110,
11110200000020,	10010020101112,	10100310110101,	11000000110101,
01101201100023,	01010201111121,	00110121001113,	11011120001022,
10000000010103,	11011310001001,	10111300001031,	01011211111020,
10001130010010,	01100301100131,	00001231100013,	10100200000121,
01010321111133,	00111001111030,	11110010000001,	01111201000120,
01001321011132,	11011220010122,	00111331001032,	01010301001100,
01101301100033,	00000101111010,	00011221101032,	00101231110033,
10111100111020,	11000010110102,	11010300001120,	01010021111103,
10010330011130,	01010001010000,	10010100101120,	11101320001132,
11100200100123,	11000000000110,	01010131111110,	01001121000022,
11101330001133,	00000131100123,	00111211001020,	11100120100111,
00001131100003,	00011321101002,	00101211000022,	10111010100101,
01011021001032,	01100301010100,	10011100101000,	00101001000001,
01000301000120,	01011311001021,	10000210010120,	00101331110003,
00000001100110,	10111020111012,	00110021001103,	00110121111122,
10100300000131,	01111031011013,	00100101110100,	01100331010103,
10111230111033,	11100110100110,	10100110000112,	10001230100033,
11011020001012,	10011230011022,	11000200110121,	10101200000023,
10101120000011,	11111200101023,	01001201011120,	10100320110102,
11100020100101,	00101131000010,	00110021111112,	11100220010112,
00000101010111,	00001131010012,	01100001001000,	00000231111023,
10110100111100,	10010310110031,	00000121100122,	10101300000033,
10000000100130,	01010221001132,	01010131010013,	01100321001032,
10010010011102,	00011111101021,	11001020110001,	01101121100011,
01100021010112,	10000100100100,	11011320001002,	00110321111102,
00101211101121,	10011300011033,	00000021111002,	01010321010032,
01001231011123,	11011110111012,	11111030011033,	10111300100130,
10010110101121,	10100130110123,	00000331111033,	01111221000122,
01011121111011,	10011010011000,	00100311000130,	00101221000023,
00010331101123,	11101220100023,	01110111011101,	10100220110132,
11101010010011,	00101121000013,	01001221000032,	10000310010130,
01011231111022,	10010230110023,	11011310111032,	01100001100101,
00110131100013,	10001030111103,	00100321000131,	00011021011003,
00000111100121,	01111331011003,	11010200001110,	01101331100032,
00111231001022,	00110211111131,	10010230101133,	11110320101133,
00011311110131,	10111100100110,	11101010001101,	01100121001012,
01001011000011,	11011120010112,	00001331100023,	00100231000122,
10001100100020,	11101030010013,	00010101011113,	00110231100023,
01000201110123,	01001231000033,	00000221100132,	10110110001110,
01000331110132,	00000001111000,	01011221111021,	11000320011032,

10111220100122,	11101210100022,	11010220111121,	11010030111102,
01110131101112,	11000210110122,	10010030101113,	10010020110002,
01100201001020,	10000200010123,	11010230001113,	11010230111122,
00011031110103,	10100220000123,	11111110101010,	00100121000111,
00011101110110,	10111310100131,	10111120111022,	00111131001012,
11010130111112,	11111210011011,	00011031011000,	10010110110011,
10111020001003,	10011200011023,	10111330111003,	00000211111021,
01111121011022,	01101131010003,	00111121001011,	00100301110120,
11010000111103,	01011221001012,	00010211101111,	01000211000111,
11101300001130,	01001211000031,	01101111010001,	01000231000113,
00101331101133,	00101211110031,	01100331001033,	01011111001001,
10011330101023,	10011310011030,	00010201101110,	11010000001130,
11000100011010,	11011230111020,	10111310001032,	00001221100012,
11011200001030,	10100020101002,	10100210000122,	10011020011001,
00000311100101,	11001120000002,	11010010111100,	01101111100010,
00111321001031,	10101210110011,	11110210011131,	00001201010023,
10111320001033,	11010310111130,	11100000010130,	01001221110023,
10011230101013,	11110030000003,	00110331100033,	00111211111011,
11111000101003,	01010331010033,	01101221010012,	00101321110002,
00010121011111,	11001200110023,	01001021000012,	11000330000103,
10100010101001,	11100300010120,	01100321100133,	00011221011023,
01001211011121,	11011000010100,	11011130111010,	11110330011103,
00011121101022,	11010120111111,	11110020101103,	10111130111023,
11101020010012,	01011301111033,	10001210100031,	11001200000010,
10000110010110,	11000220110123,	11011020010102,	01000121000102,
01000221000112,	00000011100111,	01000001110103,	11110100011120,
11100310100130,	00110031001100,	00101131110023,	01110231101122,
00100311110121,	11000110011011,	11100100100113,	00000301010131,
10100100110120,	11110010011111,	10001110111111,	10011220101012,
11011300111031,	01100301001030,	00110011100001,	01010221010022,
10111330001030,	00011021101012,	10100200110130,	10100020110112,
01011001001030,	00000211010122,	00001121100002,	00110201001121,
10111330100133,	01010101001120,	01110001101103,	00000111111011,
01100311100132,	10110200111110,	01110131011103,	00011231101033,
01100311010101,	11000310110132,	00111201001023,	00000201010121,
01011301001020,	01100031001003,	10001120010013,	10101330000032,
01101311010021,	10111000100100,	10110130001112,	10101310110021,
00001031010002,	01110221101121,	00001101010013,	01110211011111,
10010130110013,	10010010110001,	10010000101110,	10100310101031,
01111131101010,	01001121011112,	10100320000133,	11011110010111,
00111221111012,	11001130110012,	11111300101033,	11011330001003,
10000320010131,	00101031101103,	11000220011022,	00100331110123,
10011120101002,	00100221000121,	11111330011023,	11011210001031,
11011010001011,	01100131100110,	11111120101011,	00011011011002,
01010211010021,	11110000011110,	00101111110021,	01001111000021,
00100301000133,	01001331110030,	01000321110131,	00001011010000,
00100201110110,	11000020011002,	11010100001100,	11011220001032,
00110101001111,	10010020011103,	00111311001030,	11100100010100

Appendix 10: Equity Weights for Economic Evaluation DCE

Page 1: Welcome




Welcome

We are inviting you to participate in a study designed to gain an understanding of people's opinions about the importance of health to different groups in society.


This is part of a PhD project, and the answers you provide are important in investigating the topic. The information you provide will help decision makers in Australia to focus on the areas that Australians value highest. The study is being undertaken by the Centre for Health Economics Research and Evaluation at the University of Technology, Sydney (UTS).

Your participation in this project is voluntary, and you may drop out at any stage.

progress 

[next](#)

Page 2: Survey Structure




The survey contains three sections.

Section A contains sixteen questions. In each question you will be shown two hypothetical health programs. Each health program would improve the life expectancy of a group of one hundred individuals with particular characteristics. These programs are not intended to represent existing options and have been made up for the purpose of this exercise. Each option is described in terms of the characteristics of the group who might receive the program, and the gain in life expectancy that the program would give the group if it were implemented. Your task will be to choose which option you would choose if you could only choose one of them.

Section B contains questions about you, like your age, which will allow us to apply the results of this study to the population as a whole.

Section C is a brief feedback form, identifying where this survey could be improved.

progress 

[prev](#) [next](#)

Page 3: Introduction to Task



Section A: Making Choices Between Options

You will now be presented with sixteen choices. In each situation, you will see two health programs.

Consider the example below. The text under the heading 'program One' describes a health program which might be given to one hundred individuals (called Group 1). These individuals have the characteristics listed. program one is to be compared with program two, which is a health program which might be given to one hundred individuals (called Group 2).

If you were asked to choose one of the following two Programs, each of which would impact on the health of 100 people, which would you select?

	Program One	Program Two
The people in this group are	Female	Male
The people in this group have an income which is	Below average	Above average
The people in this group are	Non-smokers	Non-smokers
In terms of diet, exercise and avoiding high risk activities, the people in this group have a lifestyle which is generally	Healthy	Healthy
Are the people in this groups fulltime carers (e.g. for children)	Yes	No
Without the program, the people in the group will live until they are	60 years	75 years
The program would increase their life expectancy by	3 years	3 years
Which program would you choose?	<input type="radio"/>	<input type="radio"/>

Your task is to consider the characteristics of the two groups, and the potential health gain they would experience under the health programs. Then, you decide which of the two health programs you would pick if you had to choose between them. There are no right or wrong answers. Some of the programs may be difficult for you to imagine - just do the best you can.

progress

[prev](#) [next](#)

Pages 4-19: Choice Sets 1-16




If you were asked to choose one of the following two programs, each of which would impact on the health of 100 people, which would you select?

	Program One	Program Two
The people in this group are	Male	Female
The people in this group have an income which is	Above average	Above average
The people in this group are	Smokers	Non-smokers
In terms of diet and exercise the people in this group have a lifestyle which is generally	Healthy	Healthy
Are the people in this group fulltime carers (e.g. for children or for adults with medical conditions)?	Yes	No
Without the program, the people in the group will live until they are	30 years	30 years
The program would increase their life expectancy by	1 year	10 years
Which program would you choose?	<input type="radio"/>	<input type="radio"/>

progress

[prev](#) [next](#)


Page 20: Response Approach



You have now finished the 16 questions. Thank you for your contribution so far. Before you finish, we need to collect some information about you. This helps us to ensure we have responses from all types of people.


In general, which of this statements best represents your approach to answering the task?

- I considered all of the areas of health and life expectancy before making my decision.
- I focused on some of the areas and made my decision based on these.
- I selected at random.

progress 

[prev](#) [next](#)

Page 21: Demographics I



Section B: Some more information about you


Are you:

- Male
- Female

How old are you (in years)?


Are you

- Aboriginal
- Torres Strait Islander
- Neither
- Both

progress 


[prev](#) [next](#)

Page 22: Demographics II




Country of Birth

- Australia
- England
- New Zealand
- Italy
- Vietnam
- Scotland
- Greece
- Germany
- Philippines
- Netherlands
- Other (please specify)

progress 

[prev](#) [next](#)

Page 23: Demographics III



What is the highest level of education you completed?


- Completed primary
- Completed secondary
- Trade certificate/Diploma
- Bachelor's degree
- Higher degree

Are you currently studying?


- Yes
- No

Please indicate which of these categories best matches your before tax household income (you may refuse to answer if you wish).

- Under \$20,000
- \$20,001-\$30,000
- \$30,001-\$40,000
- \$40,001-\$50,000
- \$50,001-\$60,000
- \$60,001-\$70,000
- \$70,001-\$80,000
- \$80,001-\$100,000
- Over \$100,000

progress 

[prev](#) [next](#)



What is your current marital status? (Please skip if you would rather not answer this)


Single
 Separated/Divorced
 Widowed
 Married/De facto

Please enter the number of children for each of the age ranges below


	number of children
less than 5 years	<input type="text"/>
aged between 5 and 16 years	<input type="text"/>
older than 16 years?	<input type="text"/>

Are you a carer (e.g. for children or for adults with medical conditions)?

Yes
 No

progress 

[prev](#) [next](#)




What is your smoking status?

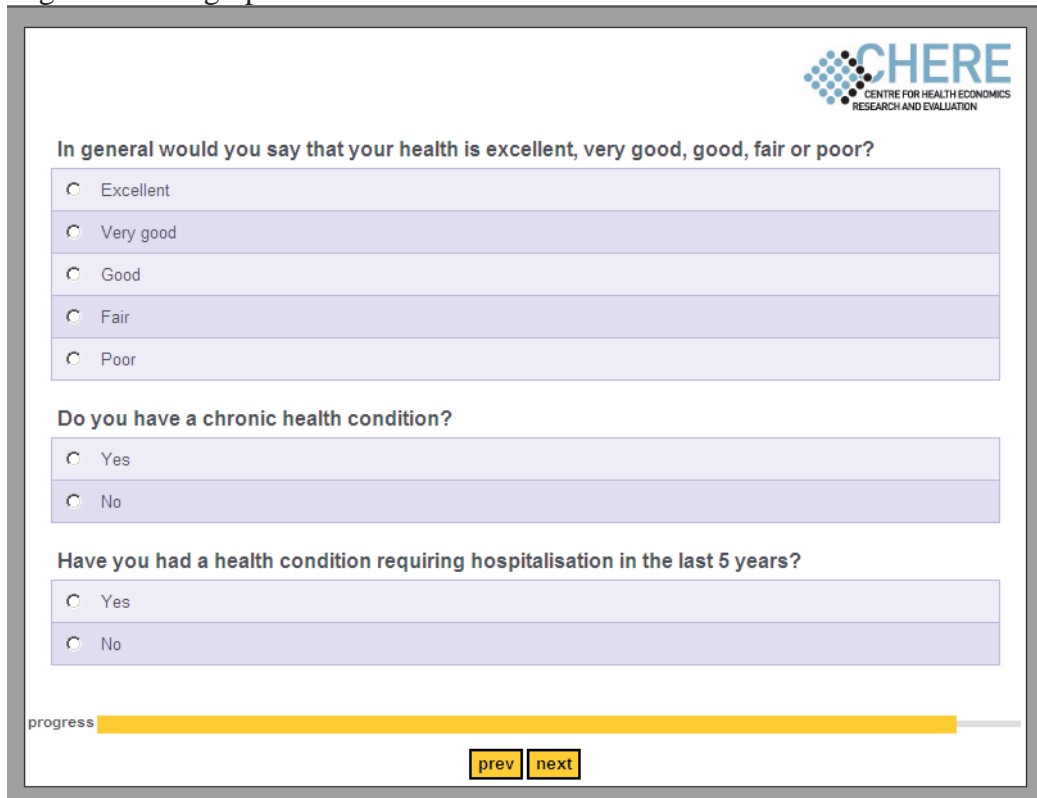
I am a current smoker
 I am a former smoker
 I have never regularly smoked

In terms of diet and exercise you have a lifestyle which is generally healthy

Strongly agree
 Agree
 Neutral
 Disagree
 Strongly disagree

progress 

[prev](#) [next](#)



CHERE
CENTRE FOR HEALTH ECONOMICS
RESEARCH AND EVALUATION

In general would you say that your health is excellent, very good, good, fair or poor?

Excellent

Very good

Good

Fair

Poor

Do you have a chronic health condition?


Yes

No

Have you had a health condition requiring hospitalisation in the last 5 years?

Yes

No

progress 

[prev](#) [next](#)

Section C: Improving the Survey

Thankyou for completing the survey. We would like to ask you to answer the following question, identifying how easy you found the survey to answer. If you wish to add any details of this, please add this in the box below. Thankyou in advance for any feedback you may have.

In general, how clear was the task that you were asked to complete (i.e. how easy was it to understand the question?)?

- Very clear
- Quite clear
- Neither clear nor unclear
- Quite unclear
- Very unclear

In general, how difficult was the task that you were asked to complete?

- Very difficult
- Quite difficult
- Neither difficult nor easy
- Quite easy
- Very easy

If you have any general comments about the survey, please give them here

progress 

[prev](#) [next](#)

Appendix 11: Equity Weights gender subgroup analysis

```
. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthyife
gain_iscarer gain_le45 gain_le60 gain_le75 if xfemale==1
```

```
Random-effects probit regression           Number of obs       =       4960
Group variable: rid                       Number of groups    =        310

Random effects u_i ~ Gaussian              Obs per group: min =        16
                                           avg =           16.0
                                           max =           16

Log likelihood = -3050.8707                Wald chi2(9)        =       644.28
                                           Prob > chi2         =       0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gain_linear	.1111325	.0094197	11.80	0.000	.0926702 .1295948
gain_female	.0239103	.0032582	7.34	0.000	.0175244 .0302962
gain_highy	-.0130041	.0037676	-3.45	0.001	-.0203884 -.0056199
gain_smoker	-.0785796	.0045065	-17.44	0.000	-.0874123 -.069747
gain_healthyife	.0168215	.0061817	2.72	0.007	.0047056 .0289374
gain_iscarer	.0338543	.0036594	9.25	0.000	.026682 .0410267
gain_le45	.0148267	.0074287	2.00	0.046	.0002668 .0293866
gain_le60	.0134473	.008386	1.60	0.109	-.0029889 .0298836
gain_le75	-.0073072	.007515	-0.97	0.331	-.0220363 .007422
_cons	-.0418168	.0187353	-2.23	0.026	-.0785372 -.0050963
/lnsig2u	-12.80738	10.46768			-33.32366 7.708892
sigma_u	.0016554	.0086643			5.81e-08 47.20246
rho	2.74e-06	.0000287			3.37e-15 .9995514

```
Likelihood-ratio test of rho=0: chibar2(01) = 7.3e-04 Prob >= chibar2 = 0.489
```

```
. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthyife
gain_iscarer gain_le45 gain_le60 gain_le75 if xfemale==0
```

```
Random-effects probit regression           Number of obs       =       3888
Group variable: rid                       Number of groups    =        243

Random effects u_i ~ Gaussian              Obs per group: min =        16
                                           avg =           16.0
                                           max =           16

Log likelihood = -2469.0481                Wald chi2(9)        =       395.32
                                           Prob > chi2         =       0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gain_linear	.1113947	.0100268	11.11	0.000	.0917425 .131047
gain_female	-.022065	.0036625	-6.02	0.000	-.0292433 -.0148866
gain_highy	-.0025508	.004237	-0.60	0.547	-.0108553 .0057536
gain_smoker	-.0706042	.0050771	-13.91	0.000	-.0805551 -.0606533
gain_healthyife	.0133035	.0070414	1.89	0.059	-.0004973 .0271043
gain_iscarer	.0288022	.0040804	7.06	0.000	.0208048 .0367996
gain_le45	.0164388	.007991	2.06	0.040	.0007768 .0321008
gain_le60	.0046851	.0092158	0.51	0.611	-.0133775 .0227476
gain_le75	-.0130698	.0081605	-1.60	0.109	-.0290642 .0029246
_cons	-.0267959	.0208508	-1.29	0.199	-.0676627 .0140708
/lnsig2u	-12.39348	17.55597			-46.80255 22.01559

```

sigma_u | .0020361 .0178725 6.87e-11 60342.57
rho | 4.15e-06 .0000728 4.72e-21 1

```

Likelihood-ratio test of rho=0: chibar2(01) = 2.0e-04 Prob >= chibar2 = 0.494

*** Equity Weights gender subgroup analysis ***

```

. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthyife
gain_iscarer gain_le45 gain_le60 gain_le75 if xfemale==1

```

```

Random-effects probit regression      Number of obs      =      4960
Group variable: rid                  Number of groups   =      310

```

```

Random effects u_i ~ Gaussian        Obs per group: min =      16
                                      avg =      16.0
                                      max =      16

```

```

Log likelihood = -3050.8707          Wald chi2(9)      =      644.28
                                      Prob > chi2       =      0.0000

```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gain_linear	.1111325	.0094197	11.80	0.000	.0926702 .1295948
gain_female	.0239103	.0032582	7.34	0.000	.0175244 .0302962
gain_highy	-.0130041	.0037676	-3.45	0.001	-.0203884 -.0056199
gain_smoker	-.0785796	.0045065	-17.44	0.000	-.0874123 -.069747
gain_healt~e	.0168215	.0061817	2.72	0.007	.0047056 .0289374
gain_iscarer	.0338543	.0036594	9.25	0.000	.026682 .0410267
gain_le45	.0148267	.0074287	2.00	0.046	.0002668 .0293866
gain_le60	.0134473	.008386	1.60	0.109	-.0029889 .0298836
gain_le75	-.0073072	.007515	-0.97	0.331	-.0220363 .007422
_cons	-.0418168	.0187353	-2.23	0.026	-.0785372 -.0050963
/lnsig2u	-12.80738	10.46768			-33.32366 7.708892
sigma_u	.0016554	.0086643			5.81e-08 47.20246
rho	2.74e-06	.0000287			3.37e-15 .9995514

Likelihood-ratio test of rho=0: chibar2(01) = 7.3e-04 Prob >= chibar2 = 0.489

```

. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthyife
gain_iscarer gain_le45 gain_le60 gain_le75 if xfemale==0

```

```

Random-effects probit regression      Number of obs      =      3888
Group variable: rid                  Number of groups   =      243

```

```

Random effects u_i ~ Gaussian        Obs per group: min =      16
                                      avg =      16.0
                                      max =      16

```

```

Log likelihood = -2469.0481          Wald chi2(9)      =      395.32
                                      Prob > chi2       =      0.0000

```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gain_linear	.1113947	.0100268	11.11	0.000	.0917425 .131047
gain_female	-.022065	.0036625	-6.02	0.000	-.0292433 -.0148866
gain_highy	-.0025508	.004237	-0.60	0.547	-.0108553 .0057536
gain_smoker	-.0706042	.0050771	-13.91	0.000	-.0805551 -.0606533
gain_healt~e	.0133035	.0070414	1.89	0.059	-.0004973 .0271043
gain_iscarer	.0288022	.0040804	7.06	0.000	.0208048 .0367996
gain_le45	.0164388	.007991	2.06	0.040	.0007768 .0321008
gain_le60	.0046851	.0092158	0.51	0.611	-.0133775 .0227476
gain_le75	-.0130698	.0081605	-1.60	0.109	-.0290642 .0029246
_cons	-.0267959	.0208508	-1.29	0.199	-.0676627 .0140708

```

-----+-----
/lnsig2u | -12.39348  17.55597                -46.80255  22.01559
-----+-----
sigma_u  | .0020361  .0178725                6.87e-11  60342.57
rho      | 4.15e-06  .0000728                4.72e-21  1
-----+-----
Likelihood-ratio test of rho=0: chibar2(01) = 2.0e-04 Prob >= chibar2 = 0.494

```

Appendix 12: Equity Weights smoker subgroup analysis

```
. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthy life
gain_iscarerer gain_le45 gain_le60 gain_le75 if xsmoke==1
```

```
Random-effects probit regression           Number of obs       =       1648
Group variable: rid                       Number of groups    =        103

Random effects u_i ~ Gaussian             Obs per group: min =        16
                                           avg =       16.0
                                           max =        16

Log likelihood = -1086.1773                Wald chi2(9)        =       100.83
                                           Prob > chi2         =        0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gain_linear	.0533637	.0147216	3.62	0.000	.0245098 .0822175
gain_female	-.0048422	.0053825	-0.90	0.368	-.0153918 .0057074
gain_highy	-.01023	.0062774	-1.63	0.103	-.0225334 .0020734
gain_smoker	-.0011038	.0071969	-0.15	0.878	-.0152095 .0130019
gain_healt~e	.031724	.0099927	3.17	0.001	.0121386 .0513093
gain_iscarerer	.0248583	.0060857	4.08	0.000	.0129305 .0367861
gain_le45	.0162594	.0124474	1.31	0.191	-.0081371 .0406559
gain_le60	.0218542	.0139462	1.57	0.117	-.0054799 .0491883
gain_le75	.005134	.0122681	0.42	0.676	-.018911 .029179
_cons	-.0784653	.0315832	-2.48	0.013	-.1403673 -.0165633
/lnsig2u	-15.69052	18.36266			-51.68067 20.29963
sigma_u	.0003916	.0035954			5.99e-12 25586.38
rho	1.53e-07	2.82e-06			3.59e-23 1

```
Likelihood-ratio test of rho=0: chibar2(01) = 0.00 Prob >= chibar2 = 1.000
```

```
. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthy life
gain_iscarerer gain_le45 gain_le60 gain_le75 if xsmoke==2
```

```
Random-effects probit regression           Number of obs       =       2384
Group variable: rid                       Number of groups    =        149

Random effects u_i ~ Gaussian             Obs per group: min =        16
                                           avg =       16.0
                                           max =        16

Log likelihood = -1502.9223                Wald chi2(9)        =       254.46
                                           Prob > chi2         =        0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gain_linear	.1056183	.0135427	7.80	0.000	.0790751 .1321615
gain_female	-.0046569	.004775	-0.98	0.329	-.0140156 .0047019
gain_highy	-.0020351	.0055389	-0.37	0.713	-.0128911 .0088209
gain_smoker	-.0778522	.006557	-11.87	0.000	-.0907037 -.0650006
gain_healt~e	.0167515	.0090112	1.86	0.063	-.0009101 .034413
gain_iscarerer	.0308998	.0053467	5.78	0.000	.0204205 .0413791
gain_le45	.0182201	.011002	1.66	0.098	-.0033434 .0397837
gain_le60	.003862	.0122108	0.32	0.752	-.0200707 .0277947
gain_le75	-.0048195	.0106837	-0.45	0.652	-.0257592 .0161201
_cons	-.0283684	.0272208	-1.04	0.297	-.0817202 .0249835
/lnsig2u	-5.637662	3.685329			-12.86077 1.585451
sigma_u	.0596757	.1099623			.0016118 2.20941
rho	.0035485	.0130312			2.60e-06 .8299751

Appendix 13: Equity Weights carer status subgroup analysis

```
. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthyife
gain_iscarer gain_le45 gain_le60 gain_le75 if xcarer==1
```

```
Random-effects probit regression           Number of obs       =       1744
Group variable: rid                       Number of groups    =        109

Random effects u_i ~ Gaussian             Obs per group: min =        16
                                           avg =       16.0
                                           max =        16

Log likelihood = -1088.2375                Wald chi2(9)        =       208.34
                                           Prob > chi2         =       0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
gain_linear	.095296	.0157386	6.05	0.000	.0644489	.1261431
gain_female	.0092815	.0054569	1.70	0.089	-.0014137	.0199768
gain_highy	-.0120714	.0064099	-1.88	0.060	-.0246345	.0004917
gain_smoker	-.0738925	.0073305	-10.08	0.000	-.0882601	-.0595248
gain_healt~e	.0042248	.0102024	0.41	0.679	-.0157716	.0242211
gain_iscarer	.0489755	.006133	7.99	0.000	.036955	.060996
gain_le45	.0153945	.0127333	1.21	0.227	-.0095623	.0403514
gain_le60	.0093859	.0141401	0.66	0.507	-.0183281	.0371
gain_le75	-.0107986	.0128213	-0.84	0.400	-.0359279	.0143307
_cons	-.0551244	.03152	-1.75	0.080	-.1169025	.0066536
/lnsig2u	-14.11102	23.8034			-60.76482	32.54279
sigma_u	.0008626	.0102669			6.38e-14	1.17e+07
rho	7.44e-07	.0000177			4.08e-27	1

```
Likelihood-ratio test of rho=0: chibar2(01) = 0.00 Prob >= chibar2 = 1.000
```

```
. xtprobit choice gain_linear gain_female gain_highy gain_smoker gain_healthyife
gain_iscarer gain_le45 gain_le60 gain_le75 if xcarer==2
```

```
Random-effects probit regression           Number of obs       =       7088
Group variable: rid                       Number of groups    =        443

Random effects u_i ~ Gaussian             Obs per group: min =        16
                                           avg =       16.0
                                           max =        16

Log likelihood = -4459.2897                Wald chi2(9)        =       778.35
                                           Prob > chi2         =       0.0000
```

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
gain_linear	.1158256	.0076241	15.19	0.000	.1008827	.1307685
gain_female	.0010243	.0027105	0.38	0.705	-.0042882	.0063369
gain_highy	-.0085474	.0031335	-2.73	0.006	-.014689	-.0024059
gain_smoker	-.0743339	.0037643	-19.75	0.000	-.0817118	-.066956
gain_healt~e	.0183744	.0051797	3.55	0.000	.0082223	.0285264
gain_iscarer	.0287044	.0030414	9.44	0.000	.0227434	.0346655
gain_le45	.009064	.0061107	1.48	0.138	-.0029129	.0210408
gain_le60	.0072691	.0069074	1.05	0.293	-.0062692	.0208074
gain_le75	-.0108814	.0061109	-1.78	0.075	-.0228585	.0010957
_cons	-.0300547	.0155115	-1.94	0.053	-.0604567	.0003473
/lnsig2u	-12.46204	9.573827			-31.2264	6.302312
sigma_u	.0019674	.009418			1.66e-07	23.36306
rho	3.87e-06	.0000371			2.75e-14	.9981713

Likelihood-ratio test of rho=0: chibar2(01) = 9.8e-04 Prob >= chibar2 = 0.487

Appendix 14: Variance Covariance Matrices (Utility Function A)

Variance Covariance Matrix for Model A5

Mean (se)	Gain	Female	High Income	Smoker	Healthy Lifestyle	Carer	Life expect 45	Life expect 60	Life expect 75
Gain	0.701 (0.053) ***								
Female	-0.014 (0.020)	0.267 (0.017) ***							
High Income	-0.024 (0.020)	-0.054 (0.019) ***	-0.212 (0.018) ***						
Smoker	-0.176 (0.042) ***	-0.085 (0.029) ***	0.008 (0.039)	-0.437 (0.030) ***					
Healthy Lifestyle	-0.052 (0.030) *	0.023 (0.027)	0.034 (0.028)	0.062 (0.037) *	0.211 (0.029) ***				
Carer	0.045 (0.019) **	-0.016 (0.022)	0.089 (0.022) ***	0.027 (0.020)	-0.018 (0.022)	-0.269 (0.019) ***			
Life expect 45	-0.200 (0.033) ***	0.006 (0.032)	-0.028 (0.033)	0.042 (0.030)	0.038 (0.033)	0.137 (0.030) ***	0.208 (0.030) ***		
Life expect 60	-0.344 (0.040) ***	-0.029 (0.035)	0.022 (0.039)	0.122 (0.036) ***	0.088 (0.036) **	0.158 (0.038) ***	0.242 (0.034) ***	0.173 (0.033) ***	
Life expect 75	-0.487 (0.046) ***	-0.109 (0.037) ***	-0.030 (0.047)	0.135 (0.047) ***	0.266 (0.040) ***	0.263 (0.045) ***	0.222 (0.039) ***	0.352 (0.039) ***	0.099 (0.052) *

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

Variance Covariance Matrix for Model A6

Mean (standard error)	Gain	Female	High Income	Smoker	Healthy Lifestyle	Carer	Life expect 45	Life expect 60	Life expect 75
Gain	-1.577 (0.312) ***								
Female	0.075 (0.025) ***	0.746 (0.15) ***							
High Income	-0.225 (0.05) ***	-0.027 (0.017)	0.363 (0.071) ***						
Smoker	-0.316 (0.065) ***	-0.363 (0.091) ***	-0.346 (0.07) ***	0.743 (0.154) ***					
Healthy Lifestyle	0.345 (0.072) ***	0.273 (0.073) ***	0.163 (0.047) ***	0.32 (0.08) ***	0.167 (0.05) ***				
Carer	0.063 (0.034) *	-0.095 (0.033) ***	-0.591 (0.121) ***	-0.345 (0.076) ***	0.472 (0.093) ***	-0.257 (0.048) ***			
Life expect 45	0.283 (0.073) ***	0.022 (0.052)	0.331 (0.072) ***	0.085 (0.049) *	-0.062 (0.041)	0.32 (0.068) ***	0.596 (0.121) ***		
Life expect 60	0.788 (0.157) ***	0.1 (0.056) *	0.535 (0.122) ***	0.258 (0.088) ***	0.175 (0.055) ***	0.54 (0.105) ***	0.633 (0.13) ***	0.327 (0.069) ***	
Life expect 75	1.087 (0.22) ***	0.12 (0.057) **	0.753 (0.157) ***	0.369 (0.102) ***	0.475 (0.11) ***	1.087 (0.218) ***	0.589 (0.13) ***	0.855 (0.19) ***	-0.375 (0.089) ***

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

Appendix 15: Variance Covariance Matrices (Utility Function B)

Variance Covariance Matrix for Model B5

Mean (standard error)	Gain	Female	High Income	Smoker	Healthy Lifestyle	Carer	Life expect 45	Life expect 60	Life expect 75
Gain	0.678 (0.054) ***								
Female	-0.013 (0.021)	0.261 (0.017) ***							
High Income	-0.012 (0.020)	-0.051 (0.021) **	-0.207 (0.018) ***						
Smoker	-0.156 (0.043) ***	-0.077 (0.030) **	0.008 (0.035)	-0.427 (0.029) ***					
Healthy Lifestyle	-0.050 (0.031)	0.025 (0.029)	0.038 (0.028)	0.059 (0.035) *	0.197 (0.030) ***				
Carer	0.038 (0.019) *	-0.013 (0.024)	0.087 (0.023) ***	0.023 (0.020)	-0.020 (0.023)	-0.259 (0.019) ***			
Life expect 45	-0.207 (0.034) ***	0.002 (0.032)	-0.029 (0.032)	0.034 (0.029)	0.033 (0.033)	0.135 (0.029) ***	0.205 (0.031) ***		
Life expect 60	-0.351 (0.041) ***	-0.029 (0.034)	0.013 (0.039)	0.109 (0.034) ***	0.083 (0.036) **	0.157 (0.036) ***	0.235 (0.033) ***	0.164 (0.032) ***	
Life expect 75	-0.485 (0.046) ***	-0.106 (0.039) ***	-0.035 (0.052)	0.120 (0.044) ***	0.258 (0.039) ***	0.265 (0.043) ***	0.215 (0.038) ***	0.345 (0.038) ***	0.094 (0.057) *

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

Variance Covariance Matrix for Model B6

Mean (standard error)	Gain	Female	High Income	Smoker	Healthy Lifestyle	Carer	Life expect 45	Life expect 60	Life expect 75
Gain	-2.426 (0.434) ***								
Female	0.021 (0.035)	1.013 (0.178) ***							
High Income	-0.098 (0.041) **	-0.208 (0.046) ***	0.387 (0.071) ***						
Smoker	1.076 (0.192) ***	-0.382 (0.077) ***	0.524 (0.101) ***	-1.404 (0.250) ***					
Healthy Lifestyle	0.268 (0.084) ***	0.267 (0.066) ***	0.475 (0.106) ***	0.360 (0.075) ***	-0.320 (0.067) ***				
Carer	-0.292 (0.068) ***	-0.127 (0.039) ***	-0.510 (0.100) ***	-0.156 (0.041) ***	-0.380 (0.064) ***	0.573 (0.100) ***			
Life expect 45	0.634 (0.132) ***	-0.04 (0.065)	0.366 (0.078) ***	0.242 (0.065) ***	0.445 (0.101) ***	0.284 (0.074) ***	0.322 (0.086) ***		
Life expect 60	1.137 (0.207) ***	-0.094 (0.064)	0.564 (0.113) ***	0.588 (0.118) ***	0.712 (0.140) ***	0.288 (0.073) ***	0.275 (0.071) ***	0.729 (0.130) ***	
Life expect 75	1.768 (0.324) ***	-0.027 (0.073)	1.122 (0.203) ***	0.971 (0.184) ***	0.678 (0.124) ***	0.478 (0.111) ***	-0.153 (0.062) **	1.205 (0.203) ***	-0.379 (0.090) ***

Statistical significance noted at the 1% level (***), the 5% level (**) and the 10% level (*)

Bibliography

- Addelman S. 1961. Irregular fractions of the 2^n factorial. *Technometrics* **4**: 479-496.
- Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**: 716-723.
- Arora N, Huber J. 2001. Improving parameter estimates and model prediction by aggregate customization in choice experiments. *Journal of Consumer Research* **28**: 273-283.
- Arrow KJ. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* **53**: 941-973.
- Arrow KJ. 1950. A difficulty in the concept of social welfare. *Journal of Political Economy* **58**: 328-346.
- Arrow KJ, Lind RC. 1970. Uncertainty and the evaluation of public investment decisions. *American Economic Review* **60**: 364-378.
- Attema AE, Brouwer W. 2010. The value of correcting values: Influence and importance of correcting tto scores for time preference. *Value in Health* **13**: 879-884.
- Augustovski FA, Irazola VE, Velazquez AP, Gibbons L, Craig BM. 2009. Argentine valuation of the EQ-5D health states. *Value in Health* **12**: 587-596.
- Australian Bureau of Statistics. *Tobacco smoking in australia: A snapshot, 2004-05* Australian Bureau of Statistics: Canberra, 2006a.
- Australian Bureau of Statistics. *2006 census data by location (accessed 12th october 2011)*. Australian Bureau of Statistics: Canberra, 2006b.
- Australian Bureau of Statistics. *Population clock (accessed 12/10/11)*. Australian Bureau of Statistics: Canberra, 2011.
- Australian Bureau of Statistics. *Education and training indicators, australia, 2002*. Canberra, 2002.
- Australian Bureau of Statistics. *Population projections, australia, 2004 to 2101 (3222.0)*. Australian Bureau of Statistics: Canberra, 2005.

- Australian Bureau of Statistics. *Household income and income distribution, australia, 2005-06*. Canberra, 2007.
- Badia X, Roset M, Herdman M, Kind P. 2001. A comparison of united kingdom and spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making* **21**: 7-16.
- Badia X, Roset M, Monserrat S, Herdman M. *The spanish vas tariff based on valuation of EQ-5D health states from the general population (euroqol plenary meeting 2-3 october 1997)*. Rotterdam, 1997.
- Bala MV, Zarkin GA, Mauskopf JA. 2002. Conditions for the near equivalence of cost-effectiveness and cost-benefit analyses. *Value in Health* **5**: 338-346.
- Bansback N, Brazier J, Tsuchiya A, Anis A. 2012. Using a discrete choice experiment to estimate societal health state utility values. *Journal of Health Economics* **31**: 306-318.
- Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, et al. 2008. A comparison of the performance of the EQ-5D and SF-6D for individuals aged > 45 years. *Health Economics* **17**: 815-832.
- Becker G, Murphy K, Philipson T. *The value of life near its end and terminal care*. National Bureau of Economic Research: Cambridge, MA, 2007.
- Ben-Akiva M, Lerman S. 1985. *Discrete choice analysis: Theory and application to travel demand*. MIT: Cambridge.
- Bentham J. 1789. *An introduction to the principle of morals and legislation*. 1948 ed. Blackwell: Oxford.
- Bhat C. 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B* **35**: 677-693.
- Birch S, Donaldson C. 2003. Valuing the benefits and costs of health care programmes: Where's the 'extra' in extra-welfarism? *Social Science and Medicine* **56**: 1121-1133.
- Bleichrodt H, Diecidue E, Quiggin J. 2004. Equity weights in the allocation of health care: The rank-dependent qaly model. *Journal of Health Economics* **23**: 157-171.
- Bleichrodt H, Doctor J, Stolk E. 2005. A nonparametric elicitation of the equity-efficiency trade-off in cost-utility analysis. *Journal of Health Economics* **24**: 655-678.

- Bleichrodt H, Johannesson M. 1997. The validity of qalys: An experimental test of constant proportional tradeoff and utility independence. *Medical Decision Making* **17**: 21-32.
- Bleichrodt N, Wakker P, Johannesson M. 1997. Characterizing qalys by risk neutrality. *Journal of Risk and Uncertainty* **15**: 107-114.
- Boatwright P, Nunes JC. 2001. Reducing assortment: An attribute-based approach. *Journal of Marketing* **65**: 50-63.
- Bognar G. 2008. Age-weighting. *Economics and Philosophy* **24**: 167-189.
- Bowling A. 2005. Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health (Oxford)* **27**: 281-291.
- Box GEP, Hunter WG, Hunter JS. 1978. *Statistics for experimenters: An introduction to design, data analysis and model building*. John Wiley & Sons: New York.
- Boyle MH, Torrance GW, Sinclair JC, Horwood SP. 1983. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *New England Journal of Medicine* **308**: 1330-1337.
- Bradley RA, Terry ME. 1952. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39**: 324-345.
- Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. 2007. *Measuring and valuing health benefits for economic evaluation*. Oxford University Press: Oxford.
- Brazier J, Roberts J. 2004. The estimation of a preference-based measure of health from the sf-12. *Medical Care* **42**: 851-859.
- Brazier J, Roberts J, Deverill M. 2002. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics* **21**: 271-292.
- Brazier J, Roberts J, Tsuchiya A, Busschbach J. 2004. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics* **13**: 873-884.
- Brazier JE, Fukuhara S, Roberts J, Kharroubi S, Yamamoto Y, Ikeda S, et al. 2009. Estimating a preference-based index from the Japanese SF-36. *Journal of Clinical Epidemiology* **62**: 1323-1331.
- Brooks R, Rabin R, De Charro F. 2003. *The measurement and valuation of health status using EQ-5D: A European perspective*. Kluwer Academic Press: Dordrecht.
- Broome J. 1991. *Weighting goods*. Blackwell: Oxford.

- Brouwer WB, Culyer AJ, van Exel NJ, Rutten FF. 2008. Welfarism vs. Extra-welfarism. *Journal of Health Economics* **27**: 325-338.
- Bunch DS, Louviere J, Anderson D. *A comparison of experimental design strategies for multinomial logit models: The case of generic attributes*. University of California, Davis, 1996.
- Burgess L, Street D. 2006. The optimal size of choice sets in choice experiments. *Statistics* **40**: 507-515.
- Burgess L, Street DJ, Wasi N. In Press. Comparing designs for choice experiments using various models: A case study. *Journal of Statistical Theory and Practice*.
- Canadian Agency for Drugs and Technologies in Health. *Guidelines for the economic evaluation of health technologies: Canada*. Ottawa, 2006.
- Carlsson F, Martinsson P. 2003. Design techniques for stated preference methods in health economics. *Health Economics* **12**: 281-294.
- Chapman RG, Staelin R. 1982. Exploring rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research* **19**: 288-301.
- Charny MC, Lewis PA, Farrow SC. 1989. Choosing who shall not be treated in the nhs. *Social Science and Medicine* **28**: 1331-1338.
- Claes C, Greiner W, Uber A, Graf von der Schulenburg JM. *An interview-based comparison of the tto and vas values given to euroqol states of health by the general german population (euroqol plenary meeting 1-2 october 1998)*. Uni-Verlag Witte: University of Hannover, 1999.
- Cleemput I. *Economic evaluation in renal transplantation: Outcome assessment and cost-utility of non-compliance*. Katholieke Universiteit Leuven: Leuven, 2003.
- Coast J. 2009. Maximisation in extra-welfarism: A critique of the current position in health economics. *Social Science and Medicine* **69**: 786-792.
- Coast J, Flynn TN, Salisbury C, Louviere J, Peters TJ. 2006. Maximising responses to discrete choice experiments: A randomised trial. *Applied Health Economics and Health Policy* **5**: 249-260.
- Cochran WG, Chambers SP. 1965. The planning of observational studies of human populations. *Journal of the Royal Statistical Society (Series A (General))* **128**: 234-266.

- Culyer AJ. 1989. The normative economics of health care finance and provision. *Oxford Review of Economic Policy* **5**: 34-58.
- Culyer AJ. 1991. The normative economics of health care finance and provision. In *Providing health care: The economics of alternative systems of finance and delivery*, McGuire A, Fenn P, Mayhew K, (eds.). OUP: Oxford.
- Cystic Fibrosis Australia. *Cystic fibrosis in australia 2007 - 10th annual report from the australian cystic fibrosis data registry*. Cystic Fibrosis Australia: Sydney, 2009.
- Daniels N. 1988. *Am i my parents' keeper? An essay on justice between the young and the old*. Oxford University Press: London.
- de Bekker-Grob EW, Ryan M, Gerard K. 2012. Discrete choice experiments in health economics: A review of the literature. *Health Economics* **21**: 145-172.
- Deaton A, Muellbauer J. 1980. *Economics and consumer behaviour*. Cambridge University Press: Cambridge.
- Department of Health and Ageing. *Guidelines for preparing submissions to the pharmaceutical benefits advisory committee (version 4.2)* (<http://www.Health.Gov.Au/internet/main/publishing.Nsf/content/pbacguidelines-index>). Canberra, 2007.
- Department of Health and Ageing. *Funding for new medical technologies and procedures: Application and assessment guidelines*. Department of Health and Ageing: Canberra, 2005.
- Devlin N, Tsuchiya A, Buckingham K, Tilling C. 2011. A uniform time trade off method for states better and worse than dead: Feasibility study of the 'lead time' approach. *Health Economics* **20**: 348-361.
- Devlin NJ, Hansen P, Kind P, Williams A. 2003. Logical inconsistencies in survey respondents' health state valuations -- a methodological challenge for estimating social tariffs. *Health Economics* **12**: 529-544.
- Dey A. 1985. *Orthogonal fractional factorial designs*. Wiley: New York.
- Dolan P. 1997. Modelling valuations for euroqol health states. *Medical Care* **35**: 1095-1108.
- Dolan P. 1996. Modelling valuations for health states: The effect of duration. *Health Policy* **38**: 189-203.

- Dolan P. 2011. Thinking about it: Thoughts about health and valuing qalys. *Health Economics* **20**: 1407-1416.
- Dolan P, Cookson R, Ferguson B. 1999. Effect of discussion and deliberation on the public's views of priority setting in health care: Focus group study. *BMJ* **318**: 916-919.
- Dolan P, Gudex C, Kind P, Williams A. 1996. The time trade-off method: Results from a general population study. *Health Economics* **5**: 141-154.
- Dolan P, Gudex C, Kind P, Williams A. 1995. *A social tariff for euroqol: Results from a UK general population study. Centre for health economics york discussion paper no. 138*. Centre for Health Economics: York.
- Dolan P, Kahneman D. 2008. Interpretations of utility and their implications for the valuation of health. *The Economic Journal* **118**: 215-234.
- Dolan P, Shaw R, Tsuchiya A, Williams A. 2005. Qaly maximisation and people's preferences: A methodological review of the literature. *Health Economics* **14**: 197-208.
- Dolan P, Tsuchiya A. 2009. The social welfare function and individual responsibility: Some theoretical issues and empirical evidence. *Journal of Health Economics* **28**: 210-220.
- Donaldson C. 1998. The (near) equivalence of cost-effectiveness and cost-benefit analyses. Fact or fallacy? *Pharmacoeconomics* **13**: 389-396.
- Drukker DM, Gates R. 2006. Generating halton sequences using mata. *The STATA Journal* **6**: 214-228.
- Drummond M, O'Brien BJ, Stoddart GL, Torrance GW. 2004. *Methods for the economic evaluation of health care programmes*. Oxford Medical Publications: Oxford.
- Dworkin R. 1977. *Taking rights seriously*. Harvard University Press: Cambridge.
- Dworkin R. 1981a. What is equality? Part 2: Equality of resources. *Philosophy and Public Affairs* **10**: 283-345.
- Dworkin R. 1981b. What is equality? Part 1: Equality of welfare. *Philosophy and Public Affairs* **10**: 185-246.
- Eisenberg D, Freed GL. 2007. Reassessing how society prioritizes the health of young people. *Health Affairs* **26**: 345-354.

- El Helbawy AT, Bradley RA. 1978. Treatment contrasts in paired comparisons: Large-sample results, applications and some optimal designs. *Journal of the American Statistical Association* **73**: 831-839.
- Fanshel S, Bush J. 1970. A health status index and its application to health service outcomes. *Operations Research* **18**: 1021-1066.
- Feeny D, Furlong W, Boyle M, Torrance GW. 1995. Multi-attribute health status classification systems. Health utilities index. *Pharmacoeconomics* **7**: 490-502.
- Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. 2002. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care* **40**: 113-128.
- Feldstein M. 1963. Economic analysis, operational research, and the national health service. *Oxford Economic Papers* **15 (March)**: 19-31.
- Ferreira L, Ferreira P, Pereira L, Brazier J. 2008. An application of the SF-6D to create health values in portuguese working age adults. *Journal of Medical Economics* **11**: 215-233.
- Ferrini S, Scarpa R. 2007. Designs with a priori information for non-market valuation with choice experiments: A monte carlo study. *Journal of Environmental Economics and Management* **53**: 342-363.
- Fiebig D, Keane M, Louviere J, Wasi N. 2010. The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science* **29**: 393-421.
- Fisher I. 1918. Is "utility" the most suitable term for the concept it is used to denote? *American Economic Review* **8**: 335.
- Fisher RA. 1935. *The design of experiments*. MacMillan: Hampshire, U.K.
- Fisher RA. 1925. *Statistical methods for research workers*. Oliver and Boyd: Edinburgh.
- Flynn T. 2010. Using conjoint analysis to estimate health state values for cost-utility analysis: Issues to consider. *Pharmacoeconomics* **28**: 711-722.
- Flynn TN, Louviere JJ, Marley AA, Coast J, Peters TJ. 2008. Rescaling quality of life values from discrete choice experiments for use as qalys: A cautionary tale. *Population Health Metrics* **6**: 6.

- Flynn TN, Louviere JJ, Peters TJ, Coast J. 2010. Using discrete choice experiments to understand preferences for quality of life. Variance-scale heterogeneity matters. *Social Science and Medicine* **70**: 1957-1965.
- Gafni A. 1994. The standard gamble method: What is being measured and how it is interpreted. *Health Services Research* **29**: 207-224.
- Gan TJ, Lubarsky DA, Flood EM, Thanh T, Mauskopf J, Mayne T, et al. 2004. Patient preferences for acute pain treatment. *British Journal of Anaesthesia* **92**: 681-688.
- Gold M. 1996. *Cost-effectiveness in health and medicine*. OUP: New York.
- Golicki D, Jakubczyk M, Niewada M, Wrona W, Busschbach JJ. 2010. Valuation of EQ-5D health states in poland: First tto-based social value set in central and eastern europe. *Value in Health* **13**: 289-297.
- Gonçalves Campolina A, Bruscatto Bortoluzzo A, Bosi Ferraz M, Mesquita Ciconelli R. 2009. Validity of the SF-6D index in brazilian patients with rheumatoid arthritis. *Clinical and Experimental Rheumatology* **27**: 237-245.
- Greene W. 2003. *Econometric analysis*. 5th Edition ed. Prentice Hall: Saddle River.
- Greene WH, Hensher DA. 2011. Does scale heterogeneity across individuals matter? An empirical assessment of alternative logit models. *Transportation* **37**: 413-428.
- Greiner W, Claes C, Busschbach JJ, Graf von der Schulenburg JM. 2005. Validating the EQ-5D with time trade off for the german population. *European Journal of Health Economics* **6**: 124-130.
- Greiner W, Weijnen T, Nieuwenhuizen M, Oppe S, Badia X, Busschbach J, et al. 2003. A single european currency for EQ-5D health states. Results from a six-country study. *European Journal of Health Economics* **4**: 222-231.
- Gu Y, Hole AR, Knox S. 2011. Estimating the generalized multinomial logit model in stata (unpublished manuscript).
- Gyrd-Hansen D, Sogaard J. 2001. Analysing public preferences for cancer screening programmes. *Health Economics* **10**: 617-634.
- Haas M. 2005. The impact of non-health attributes of care on patients' choice of gp. *Australian Journal of Primary Health* **11**: 40-46.

- Hakim Z, Pathak DS. 1999. Modelling the euroqol data: A comparison of discrete choice conjoint and conditional preference modelling. *Health Economics* **8**: 103-116.
- Hall J, Fiebig DG, King MT, Hossain I, Louviere JJ. 2006a. What influences participation in genetic carrier testing? Results from a discrete choice experiment. *Journal of Health Economics* **25**: 520-537.
- Hall J, Gafni A, Birch S. *Health economics critiques of welfarism and their compatibility with sen's capabilities approach. Chere working paper series 2006/16*. Sydney, 2006b.
- Harberger AC. 1971. Three basic postulates for applied welfare economics: An interpretative essay. *Journal of Economic Literature* **9**: 785-797.
- Harris K, Keane M. 1999. A model of health plan choice: Inferring preferences and perceptions from a combination of revealed preference and attitudinal data. *Journal of Econometrics* **89**: 131-157.
- Hausman DM, McPherson MS. 1996. *Economic analysis and moral philosophy*. Cambridge University Press: Cambridge.
- Hausman J, Wise D. 1978. A conditional probit model for qualitative choice: Discrete decisions recognising interdependence and heterogeneous preferences. *Econometrica*: 403-429.
- Hawthorne G, Osborne R. 2005. Population norms and meaningful differences for the assessment of quality of life (aqol) measure. *Australian and New Zealand Journal of Public Health* **29**: 136-142.
- Hawthorne G, Richardson J, Day NA. 2001. A comparison of the assessment of quality of life (aqol) with four other generic utility instruments. *Annals of Medicine* **33**: 358-370.
- Hawthorne G, Richardson J, Day NA, Osborne R, McNeil H. *Construction and utility scaling of the assessment of quality of life (aqol) instrument*. Monash University: Melbourne, 2000.
- Hawthorne G, Richardson J, 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.
- Hensher DA, Louviere JJ, Swait J. 1999. Combining sources of preference data. *Journal of Economics* **89**: 197-221.

- Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. 2011. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5l). *Quality of Life Research* **20**: 1727-1736.
- Hicks JR. 1939. The foundations of welfare economics. *The Economic Journal* **49**: 696-712.
- Hildreth C, Houck JP. 1968. Some estimators for a model with random coefficients. *Journal of American Statistical Association* **63**: 584-595.
- Hole A. 2007a. Fitting mixed logit models by using maximum simulated likelihood. *The STATA Journal* **7**: 388-401.
- Hole AR. 2007b. A comparison of approaches to estimating confidence intervals for willingness to pay measures. *Health Economics* **16**: 827-840.
- Hole AR. 2008. Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment. *Journal of Health Economics* **27**: 1078-1094.
- Horsman J, Furlong W, Feeny D, Torrance G. 2003. The health utilities index (hui): Concepts, measurement properties and applications. *Health and Quality of Life Outcomes* **1**: 54.
- Huber J, Zwerina K. 1996. The importance of utility balance in efficient choice designs. *Journal of Marketing Research* **33**: 307-317.
- Hurley J. 2000. An overview of the normative economics of the health sector. In *Handbook of health economics*, Culyer AJ, Newhouse JP, (eds.). Elsevier: Amsterdam.
- Iyengar SS, Lepper MR. 2000. When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology* **79**: 995-1006.
- Janssen MF, Birnie E, Bonsel GJ. 2008a. Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. *Quality of Life Research* **17**: 463-473.
- Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. 2008b. Comparing the standard EQ-5D three-level system with a five-level version. *Value in Health* **11**: 275-284.
- Jelsma J, Hansen K, De Weerd W, De Cock P, Kind P. 2003. How do zimbabweans value health states? *Population Health Metrics* **1**: 11.

- Jo MW, Yun SC, Lee SI. 2008. Estimating quality weights for EQ-5D health states with the time trade-off method in south korea. *Value in Health* **11**: 1186-1189.
- Johannesson M, Johannsson PO. 1997. Is the valuation of a qaly gained independent of age? Some empirical evidence. *Journal of Health Economics* **16**: 589-599.
- Kahneman D, Tversky A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* **47**: 263-291.
- Kahneman D, Tversky A. 1982. The psychology of preferences. *Scientific American Journal I*: 160-173.
- Kanninen B. 2002. Optimal designs for multinomial choice experiments. *Journal of Marketing Research* **39**: 214-217.
- Kant I. 1785 (translation 1959). *Foundations of the metaphysics of morals (translated by lewis white beck)*. Library of Liberal Arts:
- Kaplan RM, Tally S, Hays RD, Feeny D, Ganiats TG, Palta M, et al. 2011. Five preference-based indexes in cataract and heart failure patients were not equally responsive to change. *Journal of Clinical Epidemiology* **64**: 497-506.
- Keane M, Wasi N. *Comparing alternative models of heterogeneity in consumer choice behavior*. UNSW Working Paper Series, available at http://research.economics.unsw.edu.au/mkeane/MM-MNL_June_15_09.pdf: Sydney, 2009.
- Keeney R, Raiffa H. 1993. *Decisions with multiple objectives: Preferences and value tradeoffs*. 2nd ed. Cambridge University Press: New York.
- Kessels R, Goos P, Vandebroek M. 2006. A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research* **43**: 409-419.
- Khan MA, Richardson J. *A comparison of 7 instruments in a small, general population. Research paper 60*. Centre for Health Economics, Monash University: Melbourne, 2011.
- Kharroubi S, O'Hagan A, Brazier J. 2005. Estimating utilities from individual health state preference data: A nonparametric bayesian approach. *Applied Statistics* **54**: 879-895.
- Kharroubi SA, Brazier JE, Roberts J, O'Hagan A. 2007. Modelling SF-6D health state preference data using a nonparametric bayesian method. *Journal of Health Economics* **26**: 597-612.

- Kharroubi SA, O'Hagan A, Brazier JE. 2010. A comparison of united states and united kingdom EQ-5D health states valuations using a nonparametric bayesian method. *Statistics in Medicine* **29**: 1622-1634.
- Klarman H, Francis J, Rosenthal G. 1968. Cost-effectiveness analysis applied to the treatment of chronic renal disease. *Medical Care* **6**: 48-54.
- Konerding U, Moock J, Kohlmann T. 2009. The classification systems of the EQ-5D, the hui ii and the SF-6D: What do they have in common. *Quality of Life Research* **18**: 1249-1261.
- Kuhfeld WF. 2010. *Marketing research methods in sas: Experimental design, choice, conjoint, and graphical techniques (mr-2010)*. SAS Institute Inc.: Cary, NC, USA.
- Lam CL, Brazier J, McGhee SM. 2008. Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a chinese population. *Value in Health* **11**: 295-303.
- Lamers LM, Bouwmans CAM, van Straten A, Donker MCH, Hakkaart L. 2006a. Comparison of EQ-5D and SF-6D utilities in mental health patients. *Health Economics* **15**: 1229-1236.
- Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. 2006b. The dutch tariff: Results and arguments for an effective design for national EQ-5D valuation studies. *Health Economics* **15**: 1121-1132.
- Lancaster K. 1966. A new approach to consumer theory. *Journal of Political Economy* **74**: 132-157.
- Lancsar E, Louviere J, Flynn T. 2007. Several methods to investigate relative attribute impact in stated preference experiments. *Social Science and Medicine* **64**: 1738-1753.
- Lancsar E, Savage E. 2004. Deriving welfare measures from discrete choice experiments: Inconsistency between current methods and random utility and welfare theory. *Health Economics* **13**: 901-907.
- Lancsar E, Wildman J, Donaldson C, Ryan M, Baker R. 2011. Deriving distributional weights for qalys through discrete choice experiments. *Journal of Health Economics* **30**: 466-478.
- Layard R. 1972. *Cost-benefit analysis*. Penguin: Harmondsworth, Middlesex.

- Le Galés C, Buron C, Costet N, Rosman S, Slama PR. 2002. Development of a preference-weighted health status classification system in France: The health utilities index 3. *Health Care Management Science* **5**: 41-51.
- Le Grand J. 1991. *Equity and choice: An essay in economics and applied philosophy*. Harper Collins: London.
- Loewenstein G, Angner E. 2003. Predicting and indulging changing preferences. In *Time and decision: Economic and psychological perspectives on intertemporal choice*. Russell Sage Foundation.
- Loomes G, McKenzie L. 1989. The use of QALYs in health care decision making. *Social Science and Medicine* **28**: 299-308.
- Louvière J, Hensher DA, Swait JD. 2000. *Stated choice methods: Analysis and applications*. Cambridge University Press: New York.
- Louvière J, Street D, Burgess L. 2003. A 20+ years retrospective on choice experiments. In *Marketing research and modeling: Progress and prospects*, Wind Y, Green PE, (eds.). Kluwer: New York.
- Louvière J, Street D, Burgess L, Wasi N, Islam T, Marley AA. 2008. Modeling the choices of individuals decision makers by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modeling* **1**: 128-163.
- Louvière JJ, Carson RT, Ainslie A, Cameron TA, DeShazo JR, Hensher DA, et al. 2002. Dissecting the random component of utility. *Marketing Letters* **13**: 177-193.
- Louvière JJ, Meyer RJ, Bunch DS, Carson R, Dellaert B, Hanemann WM, et al. 1999. Combining sources of preference data for modelling complex decision processes. *Marketing Letters* **10**: 205-217.
- Mæstad O, Frithjof Norheim O. 2009. Eliciting people's preferences for the distribution of health: A procedure for a more precise estimation of distributional weights. *Journal of Health Economics* **28**: 570-577.
- Marschak J. 1960. Binary choice constraints on random utility indicators. In *Stanford symposium on mathematical methods in the social sciences*, Arrow KJ, (ed.). Stanford University Press: Stanford.
- McCabe C, Brazier J, Gilks P, Tsuchiya A, Roberts J, O'Hagan A, et al. 2006. Using rank data to estimate health state utility models. *Journal of Health Economics* **25**: 418-431.

- McCabe C, Stevens K, Roberts J, Brazier J. 2005. Health state values for the hui 2 descriptive system: Results from a UK survey. *Health Economics* **14**: 231-244.
- McFadden D. 1981. Econometric models of probabilistic choice. In *Structural analysis of discrete data with economic applications*, Manski C, McFadden D, (eds.). MIT Press: Boston.
- McFadden D. 1974. Conditional logit analysis of qualitative choice behaviour. In *Frontiers in econometrics*, Zarembka P, (ed.). New York Academic Press: New York.
- McFadden D, Train K. 2000. Mixed mnl models for discrete response. *Journal of Applied Econometrics* **15**: 447-470.
- McIntosh E, Ryan M. 2002. Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: Implications for discontinuous preferences. *Journal of Economic Psychology* **23**: 367-382.
- McKie J, Richardson J. 2003. The rule of rescue. *Social Science and Medicine* **56**: 2407-2419.
- McNeil BJ, Weichselbaum R, Pauker SG. 1978. Fallacy of the five-year survival in lung cancer. *New England Journal of Medicine* **299**: 1397-1401.
- McTaggart-Cowan H, Tsuchiya A, O'Cathain AB, J. 2011. Understanding the effect of disease adaptation information on general population values for hypothetical health states. *Social Science & Medicine* **72**: 1904-1912.
- Montgomery DC. 2005. *Design and analysis of experiments*. 6th ed. John Wiley & Sons: Hoboken, NJ.
- Mooney G. 2009. *Challenging health economics*. OUP: Oxford.
- Mooney G. 1998. "Communitarian claims" as an ethical basis for allocating health care resources. *Social Science & Medicine* **47**: 1171-1180.
- Mooney G. 2005. Communitarian claims and community capabilities: Furthering priority setting? *Social Science & Medicine* **60**: 247-255.
- Mooney G, Hall J, Donaldson C, Gerard K. 1991. Utilisation as a measure of equity: Weighing heat? *Journal of Health Economics* **10**: 475-480.
- Mooney G, Jan S, Wiseman V. 1995. Examining preferences for allocating health care gains. *Health Care Analysis* **3**: 261-265.

- Mooney G, Russell E. 2003. Equity in health care: The need for a new economics paradigm? In *Advances in health economics*, Scott A, Maynard A, Elliott R, (eds.). John Wiley & Sons: Chichester.
- Murray C, Lopez A. 1994. *Global comparative assessments in the health sector: Disease burden, expenditures and intervention packages*. World Health Organisation (WHO): Geneva.
- MVH Group. *The measurement and valuation of health. Final report on the modelling of valuation tariffs*. York Centre for Health Economics, 1995.
- National Institute for Health and Clinical Excellence. *Social value judgements: Principles for the development of nice guidance (2nd edition)*. NICE: London, 2008.
- National Institute for Health and Clinical Excellence. *The guidelines manual*. National Institute for Health and Clinical Excellence (available from: www.nice.org.uk): London, 2007.
- Nord E, Pinto JL, Richardson J, Menzel P, Ubel P. 1999. Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics* **8**: 25-39.
- Nord E, Street A, Richardson J, Kuhse H, Singer P. 1996. The significance of age and duration of effect in social evaluation of health care. *Health Care Analysis* **4**: 103-111.
- Norman R, Cronin P, Viney R, King M, Street D, Ratcliffe J. 2009. International comparisons in valuing EQ-5D health states: A review and analysis. *Value in Health* **12**: 1194-1200.
- Norman R, Gallego G. 2008. *Equity weights for economic evaluation: An australian discrete choice experiment, chere working paper 2008/5*. CHERE: Sydney.
- Norman R, King M, Clarke D, Viney R, Cronin P, Street D. 2010. Does mode of administration matter? Comparison of on line and face-to-face administration of a time trade-off task. *Quality of Life Research* **19**: 499-508.
- Norman R, Viney R. *The effect of discounting on quality of life valuation using the time trade-off, chere working paper series 2008/3*. CHERE: Sydney, 2008.
- Ohinmaa A, Eija H, Sintonen H. 1996. Modelling euroqol values of finnish adult population. In *Euroqol plenary meeting*, Badia X, Herdman M, Segura A, (eds.). Institut Universitari de Salut Publica de Catalunya: Barcelona.

- Olsen JA, Richardson J, Dolan P, Menzel P. 2003. The moral relevance of personal characteristics in setting health care priorities. *Social Science and Medicine* **57**: 1163-1172.
- Osman LM, McKenzie L, Cairns J, Friend JA, Godden DJ, Legge JS, et al. 2001. Patient weighting of importance of asthma symptoms. *Thorax* **56**: 138-142.
- Packer A. 1968. Applying cost-effectiveness concepts to the community health system. *Operations Research* **16**: 227-253.
- Patrick D, Bush J, Chen M. 1973. Methods for measuring levels of well-being for a health status index. *Health Services Research* **8**: 228-245.
- Patrick D, Erickson P. 1993. Health status and health policy: Quality of life. In *Health care evaluation and resource allocation*. Oxford University Press: New York.
- Phelps CE, Mushlin A. 1991. On the (near) equivalence of cost-effectiveness and cost-benefit analysis. *International Journal of Technology Assessment in Health Care* **7**: 12-21.
- Plackett RL. 1946. Some generalizations in the multifactorial design. *Biometrika* **33**: 328-332.
- Pliskin J, Shepard D, Weinstein M. 1980. Utility functions for life years and health status. *Operations Research* **28(1)**: 206-224.
- Cabasés JM, Gaminde I, editors. The slovenian vas tariff based on valuations of EQ-5D health states from the general population. 17th Plenary Meeting of the EuroQoL Group; 2001; Universidad Pública de Navarra.
- Propper C. 1990. Contingent valuation of time spent on nhs waiting lists. *The Economic Journal* **100**: 193-199.
- Radhakrishnan M, van Gool K, Hall J, Delatycki M, Massie J. 2008. Economic evaluation of cystic fibrosis screening: A review of the literature. *Health Policy* **85**: 133-147.
- Ratcliffe J, Brazier J, Tsuchiya A, Symonds T, Brown M. 2009. Using dce and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. *Health Economics* **18**: 1261-1276.
- Rawls J. 1999. *A theory of justice*. Harvard University Press:

- Regier DA, Ryan M, Phimister E, Marra CA. 2009. Bayesian and classical estimation of mixed logit: An application to genetic testing. *Journal of Health Economics* **28**: 598-610.
- Rice T. 1992. An alternative framework for evaluating welfare losses in the health care market. *Journal of Health Economics* **11**: 85-92.
- Rice T. 1998. *The economics of health reconsidered*. Health Administration Press: Chicago.
- Richardson J. 2002a. Empirical ethics. In *Summary measures of population health: Papers from the who global conference, marrakech, december 1999*, Murray C, Lopez A, (eds.). WHO: Geneva.
- Richardson J. 2002b. The poverty of ethical analyses in economics and the unwarranted disregard of evidence. In *Summary measures of population health: Concepts, ethics, measurement and application*, Murray CJL, Salomon JA, Mathers CD, Lopez AD, (eds.). World Health Organization: Geneva.
- Richardson J, McKie J. 2005. Empiricism, ethics and orthodox economic theory: What is the appropriate basis for decision-making in the health sector? *Social Science and Medicine* **60**: 265-275.
- Richardson J, McKie J, Bariola E. *Review and critique of health related multi attribute utility instruments. Research paper 64*. Centre for Health Economics, Monash University: Melbourne, 2011.
- Robinson A, Spencer A. 2006. Exploring challenges to tto utilities: Valuing states worse than dead. *Health Economics* **15**: 393-402.
- Rodriguez E, Pinto JL. 2000. The social value of health programmes: Is age a relevant factor? *Health Economics* **9**: 611-621.
- Ryan M. 2004. Deriving welfare measures in discrete choice experiments: A comment to lancsar and savage (1). *Health Economics* **13**: 909-912; discussion 919-924.
- Salomon JA. 2003. Reconsidering the use of rankings in the valuation of health states: A model for estimating cardinal values from ordinal data. *Population Health Metrics* **1**: 12.
- Sándor Z, Wedel M. 2005. Heterogenous conjoint choice designs. *Journal of Marketing Research* **42**: 210-218.

- Santos Silva JM. 2004. Deriving welfare measures in discrete choice experiments: A comment to lancsar and savage (2). *Health Economics* **13**: 913-918; discussion 919-924.
- Sassi F, Archard L, Le Grand J. 2001. Equity and the economic evaluation of healthcare. *Health Technology Assessment* **5**: 1-138.
- Scanlon T. 1975. Preference and urgency. *Journal of Philosophy* **72**: 655-669.
- Schkade D, Kahneman D. 1998. Does living in california make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science* **9**: 340-346.
- Schwappach DLB. 2003. Does it matter who you are or what you gain? An experimental study of preferences for resource allocation. *Health Economics* **12**: 255-267.
- Schwarz GE. 1978. Estimating the dimensions of a model. *Annals of Statistics* **6**: 461-464.
- Scitovsky T. 1941. A note on welfare propositions in economics. *Review of Economic Studies* **9**: 77-88.
- Scott A. 2001. Eliciting gps' preferences for pecuniary and non-pecuniary job characteristics. *Journal of Health Economics* **20**: 329-347.
- Sculpher M, Bryan S, Fry P, de Winter P, Payne H, Emberton M. 2004. Patients' preferences for the management of non-metastatic prostate cancer: Discrete choice experiment. *BMJ* **328**: 382.
- Sen A. 1980. Equality of what? In *The tanner lectures on human values*, McMurrin S, (ed.). Cambridge University Press: Cambridge.
- Sen A. 2009. *The idea of justice*. Penguin Books: London.
- Sen A. 1992. *Inequality re-examined*. Oxford University Press: Oxford.
- Shaw JW, Johnson JA, Coons SJ. 2005. Us valuation of the EQ-5D health states: Development and testing of the d1 valuation model. *Medical Care* **43**: 203-220.
- Sintonen H. *The 15-d measure of health related quality of life: Reliability, validity and sensitivity of its health state descriptive system*. National Centre for Health Program Evaluation: Working Paper Series No.41: Melbourne, 1994.

- Small KA, Rosen HS. 1981. Applied welfare economics with discrete choice models. *Econometrica* **49**: 105-130.
- Stiggelbout AM. 2006. Health state classification systems: How comparable are our cost-effectiveness ratios? *Medical Decision Making* **26**: 223-225.
- Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. 1994. Utility assessment in cancer patients: Adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making* **14**: 82-90.
- Stolk EA, Oppe M, Scalone L, Krabbe PFM. 2010. Discrete choice modeling for the quantification of health states: The case of the EQ-5D. *Value in Health* **13**: 1005-1013.
- Stolk EA, Pickee SJ, Ament A, Busschbach JJV. 2005. Equity in health care prioritisation: An empirical inquiry into social value. *Health Policy* **74**: 343-355.
- Street D, Burgess L, Louviere JJ. 2005. Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing* **22**: 459-470.
- Street DJ, Burgess L. 2007. *The construction of optimal stated choice experiments: Theory and methods*. Wiley: Hoboken, New Jersey.
- Streiner D, Norman G. 1995. *Health measurement scales: A practical guide to their development and use*. Oxford University Press: Oxford.
- Swait J, Louviere J. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research* **30**: 305-314.
- Szende A, Oppe M, Devlin N. 2007. *EQ-5D value sets: Inventory, comparative review and user guide*. Springer: Dordrecht, The Netherlands.
- Thurstone LL. 1927a. The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology* **21**: 384-400.
- Thurstone LL. 1927b. A law of comparative judgment. *Psychological Review* **34**: 273-286.
- Tilling C, Devlin N, Tsuchiya A, Buckingham K. 2010. Protocols for time tradeoff valuations of health states worse than dead: A literature review. *Medical Decision Making* **30**: 610-619.

- Tobin J. 1970. On limiting the domain of inequality. *Journal of Law and Economics* **13**: 263-277.
- Torrance GW. 1986. Measurement of health state utilities for economic appraisal. *Journal of Health Economics* **5**: 1-30.
- Torrance GW. 1976. Social preferences for health states: An empirical evaluation of three measurement techniques. *Socioeconomic Planning Science* **10**: 129-136.
- Torrance GW, Thomas D, Sackett D. 1972. A utility maximisation model for evaluation of health care programs. *Health Services Research* **7**: 118-133.
- Train K. 2003. *Discrete choice methods with simulation*. Cambridge University Press: Cambridge.
- Tsuchiya A. 2000. Qalys and ageism: Philosophical theories and age weighting. *Health Economics* **9**: 57-68.
- Tsuchiya A, Ikeda S, Ikegami N, Nishimura S, Sakai I, Fukuda T, et al. 2002. Estimating an EQ-5D population value set: The case of japan. *Health Economics* **11**: 341-353.
- Tsuchiya A, Miguel LS, Edlin R, Wailoo A, Dolan P. 2005. Procedural justice in public health care resource allocation. *Applied Health Economics and Health Policy* **4**: 119-127.
- Tsuchiya A, Williams A. 2005. A "fair innings" between the sexes: Are men being treated inequitably? *Social Science and Medicine* **60**: 277-286.
- Tsuchiya A, Williams A. 2001. Welfare economics and economic evaluation. In *Economic evaluation in health care: Merging theory with practice*, Drummond M, McGuire A, (eds.). Oxford University Press: Oxford.
- Varian HR. 1984. *Microeconomics analysis*. 2nd ed. Norton and Company: New York.
- Viney R, Lancsar E, Louviere J. 2002. Discrete choice experiments to measure consumer preferences for health and healthcare. *Expert Review of Pharmacoeconomics & Outcomes Research* **2**: 319-326.
- Viney R, Norman R, Brazier J, Cronin P, King MT, Ratcliffe J, et al. 2011a. An australian discrete choice experiment to value EQ-5D health states. *Unpublished Manuscript*.
- Viney R, Norman R, King MT, Cronin P, Street D, Knox S, et al. 2011b. Time trade-off derived EQ-5D weights for australia. *Value in Health* **14**: 928-936.

- Viney R, Savage E. *Health care policy evaluation: Empirical analysis of the restrictions implied by quality adjusted life years*. CHERE Working Paper Series 2006/10. Sydney, 2006.
- Viney R, Savage E, Louviere J. 2005. Empirical investigation of experimental design properties of discrete choice experiments in health care. *Health Economics* **14**: 349-362.
- von Neumann J, Morgenstern O. 1947. *Theory of games and economic behaviour*. Princeton University Press: Princeton, NJ.
- von Neumann J, Morgenstern O. 1944. *Theory of games and economic behaviour*. Princeton University Press: Princeton, NJ.
- von Winterfeldt D, Edwards W. 1986. *Decision analysis and behavioural research*. Cambridge University Press: Cambridge.
- Wagstaff A, Van Doorslaer E. 2000. Equity in health care finance and delivery. In *Handbook of health economics*, Culyer AJ, Newhouse JP, (eds.). Elsevier: Amsterdam.
- Ware J, Kosinski M, Keller S. *How to score the sf-12 physical and mental health summaries: A user's manual*. The Health Institute, New England Medical Centre, Boston MA: 1995.
- Ware J, Snow K, Kosinski M, Gandek B. *SF-36 health survey manual and interpretation guide*. The Health Institute, New England Medical Centre, Boston, MA: 1993.
- Ware JE, Kosinski M, Keller SD. 1996. A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care* **34**: 220-233.
- Ware JE, Sherbourne CD. 1992. The mos 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care* **30**: 473-483.
- Webb SR. 1968. Non-orthogonal designs of even resolution. *Technometrics* **10**: 291-300.
- Weinstein MC, Manning WG, Jr. 1997. Theoretical issues in cost-effectiveness analysis. *Journal of Health Economics* **16**: 121-128.
- Welsh M, Ramsey B, Accurso F, Cutting G. 2001. Cystic fibrosis. In *The metabolic and molecular bases of inherited disease*, Scriver C, Beaudet A, Sly W, Valle D, (eds.). McGraw-Hill: New York.

- Whitehurst DGT, Bryan S. 2011. Another study showing that two preference-based measures of health-related quality of life (EQ-5D and SF-6D) are not interchangeable. But why should we expect them to be? *Value in Health* **14**: 531-538.
- Williams A. 1997. Intergenerational equity: An exploration of the 'fair innings' argument. *Health Economics* **6**: 117-132.
- Williams A. 1985. Economics of coronary artery bypass grafting. *British Medical Journal (Clinical Research Ed.)* **291**: 326-329.
- Wilson TD, Gilbert D. 2003. Constructive and unconstructive repetitive thought. *Advances in Experimental Social Psychology* **35**: 345-411.
- Wittenberg E, Prosser LA. 2011. Ordering errors, objections and invariance in utility survey responses: A framework for understanding who, why and what to do. *Applied Health Economics and Health Policy* **9**: 225-241.
- Wittrup-Jensen KU, Lauridsen JT, Gudex C, Brooks R, Pedersen KM. 2001. Estimating danish EQ-5D tariffs using the time trade-off (tto) and visual analogue scale (vas) methods. In *Proceedings of the 18th plenary meeting of the euroqol group*, Norinder A, Pedersen KL, Roos P, (eds.). Copenhagen.
- Wooldridge J. 2003. *Introductory econometrics: A modern approach*. 2nd Edition ed. Thomson South-Western: Mason, Ohio.
- Yu J, Goos P, Vandebroek M. 2009. Efficient conjoint choice designs in the presence of respondent heterogeneity. *Marketing Science* **28**: 122-135.
- Zwerina K, Huber J, Kuhfeld WF. 2010. A general method for constructing efficient choice designs. In *Marketing research methods in sas: Experimental design, choice, conjoint, and graphical techniques (mr-2010)*. SAS Institute Inc.: Cary, NC, USA.