



The
University
Of
Sheffield.

Estimating societal preferences for the allocation of healthcare resources using stated preference methods

Chris D. Skedgel, B.A., M.D.E.

A thesis submitted for the degree of Doctor of Philosophy
Health Economics & Decision Science,
School of Health & Related Research,
The University of Sheffield

December 2013

Author's declaration

I hereby declare that this thesis is my original work, conducted and composed by me, with sources of assistance noted in the acknowledgements section. No part of this work has been previously submitted for a degree at another institution. All direct quotations are identified by quotation marks and other external sources of information are cited as appropriate.

Chris Skedgel
December 2013

Acknowledgements

I must first thank my wife, Lynn Lethbridge, for her unfailing support and the patience she demonstrated through her alternating roles of chief test pilot and thesis widow.

I am grateful to my supervisors, Allan Wailoo and Ron Akehurst, for their guidance, encouragement and insights throughout this process. They were able to keep me on track through more than six years of philosophical knots, methodological quandaries, and personal crises of confidence. I was also fortunate to benefit from the thoughtful and invariably helpful advice of Ake Tsuchiya, Arne Risa Hole, John Brazier, Monica Hernández Alava, Stuart Peacock, Yukiko Asada, Koonal Shah, Dean Regier and Nick Bansback. I would also like to acknowledge three anonymous reviewers who made helpful comments on the manuscript based on the pilot survey reported in Chapter 5. Many of these comments were incorporated into the thesis.

I want to thank Danny Rayson and Tallal Younis of the Atlantic Clinical Cancer Research Unit for their professional and personal support as I worked towards this goal, as well as Murray Brown for originally setting me on this path.

Finally, I want to thank the more than 1,500 people, including friends, family and colleagues, who contributed to different aspects of this research. This work would not have been possible without their participation.

Financial support

Financial support for this work was provided by the Canadian Centre for Applied Research in Cancer Control, the Capital Health Research Fund, and the Department of Medicine and the Atlantic Clinical Cancer Research Unit, Capital District Health Authority, Halifax, Nova Scotia. The opinions expressed within the thesis are the author's alone and do not necessarily reflect the views of the funders.

Prior presentation and publication

The empirical ethics review that makes up much of Chapter 3 was presented as a poster at the *Society for Medical Decision Making International Meeting*, October 2010.

The methods and results of the pilot survey, discussed in Chapter 5, were presented as a poster at the *Society for Medical Decision Making International Meeting*, October 2011, and published as Skedgel C, Wailoo A, Akehurst R. Choosing vs. allocating: discrete choice experiments and constant-sum paired comparisons for the elicitation of societal preferences. *Health Expectations*. 2013 [Epub ahead of print].

Portions of the primary survey were presented as podium presentations at the *Canadian Applied Research Conference in Cancer Control*, May 2012, *Priorities in Health*, September 2012, and the *Health Economists' Study Group*, June 2013.

Abstract

Most governments in the world provide some publicly funded healthcare to their citizens, but given the scarcity of resources relative to potential demand, some form of rationing or priority setting is required, and some patients must be denied effective treatment. The thesis took the position that an explicit approach based on maximising the value that society derives from healthcare is the preferred way to address this rationing problem.

Conventional health economic practice proposes that value should be equated with quality-adjusted life years (QALYs), leading to a policy of QALY maximisation, but, it is argued, not necessarily *value* maximisation. A more inclusive approach to defining value, based on societal preferences, may maximise overall well-being and be associated with greater trust and legitimacy in the priority setting process.

The thesis identified patient and program characteristics that appeared to have empirical evidence of public support as well as a defensible ethical justification in determining the strength of a patient's claim to societal healthcare resources. The relative strength of preferences for these characteristics, or the equity-efficiency trade-off, was estimated using stated preference methods. Two different methods, discrete choice experiments and constant-sum paired comparisons, were used and the response behaviours of the two elicitation methods were compared to identify a preferred method for eliciting societal preferences in the context of healthcare.

Both methods found a statistically significant equity-efficiency trade-off in an age and sex representative sample of the Canadian public as well as a convenience sample of decision-making agents. This suggested that society would be willing to sacrifice some degree of efficiency in maximising individual life year gains in order to prioritise other characteristics consistent with the promotion of equity or distributive justice in the allocation of healthcare resources. However, differences between the results of the two elicitation methods suggested some systematic procedural variance.

Contents

| | |
|--|-----------|
| Abstract | v |
| Contents | vi |
| List of tables | xi |
| List of figures | xiii |
| Acronyms | xiv |
| Chapter 1: Introduction, objectives and thesis outline | 1 |
| 1.1 Thesis objectives | 2 |
| 1.2 Thesis outline | 4 |
| 1.2.1 <i>Part one</i> | 5 |
| 1.2.2 <i>Part two</i> | 6 |
| <u>Part one: Background and review</u> | |
| Chapter 2: Normative economics and healthcare priority setting | 9 |
| 2.1 The welfarist approach | 10 |
| 2.2 The extra-welfarist approach | 11 |
| 2.3 Explicitness in priority setting | 13 |
| 2.3.1 <i>Implicit priority setting</i> | 14 |
| 2.3.2 <i>Explicit priority setting</i> | 15 |
| 2.4 Inclusiveness and objectivity within the extra-welfarist framework | 17 |
| 2.4.1 <i>The decision-maker perspective and QALY maximisation</i> | 17 |
| 2.4.2 <i>A democratic or Communitarian perspective</i> | 21 |
| 2.5 Societal preferences in priority setting: the equity-weighted QALY | 22 |
| Chapter 3: Empirical ethics review | 25 |
| 3.1 Theories of justice in the allocation of healthcare | 28 |
| 3.1.1 <i>Pure procedural theories</i> | 28 |
| 3.1.2 <i>'Side condition' principles</i> | 30 |
| 3.1.3 <i>Theories with a specific maximand</i> | 31 |
| 3.1.4 <i>Defensible theories of justice</i> | 37 |

| | |
|---|-----------|
| 3.2 Attribute Review | 38 |
| 3.2.1 Age | 40 |
| 3.2.2 Social Role & Productivity | 43 |
| 3.2.3 Lifestyle and responsibility | 44 |
| 3.2.4 Prior consumption of healthcare | 47 |
| 3.2.5 Time waited | 48 |
| 3.2.6 Societal inequality | 50 |
| 3.2.7 Desert and merit | 52 |
| 3.2.8 'Self' | 53 |
| 3.2.9 Initial severity | 53 |
| 3.2.10 Final health state | 56 |
| 3.2.11 Size of health effect | 57 |
| 3.2.12 Duration of benefit | 59 |
| 3.2.13 Direction of benefit | 61 |
| 3.2.14 Distribution of health gains | 63 |
| 3.2.15 Disease rarity | 64 |
| 3.3 Attributes in other stated preference elicitation methods | 66 |
| 3.4 Fair and relevant attributes | 69 |
| Chapter 4: Comparative review of stated preference elicitation methods | 75 |
| 4.1 Measuring preferences and choices | 76 |
| 4.1.1 The theory of value and compensatory decision making | 77 |
| 4.1.2 Random utility theory | 78 |
| 4.2 A framework for comparing stated preference methods | 79 |
| 4.3 Review of stated preference methods | 82 |
| 4.3.1 Ranking | 83 |
| 4.3.2 Conjoint ranking (best-worst scaling) | 85 |
| 4.3.3 Direct constant sum scaling | 87 |
| 4.3.4 Indirect constant sum paired comparison | 89 |
| 4.3.5 Magnitude Estimation | 91 |
| 4.3.6 Person trade-off | 93 |
| 4.3.7 Full-profile ratings | 94 |
| 4.3.8 Binary choice | 96 |
| 4.3.9 Discrete choice experiments | 98 |
| 4.4 Choosing a preferred method | 101 |
| 4.5 Other studies using CSPC or DCE methods | 105 |

Part two: Empirical work

| | |
|--|------------|
| Chapter 5: Pilot survey methods & results | 109 |
|--|------------|

| | |
|--|------------|
| 5.1 DCE and CSPC in the context of random utility theory | 110 |
| 5.2 Stated preference design | 112 |
| 5.2.1 <i>Identification of attributes</i> | 112 |
| 5.2.2 <i>Assigning levels</i> | 114 |
| 5.2.3 <i>Experimental design</i> | 116 |
| 5.2.4 <i>Stated preferences & rationality</i> | 121 |
| 5.2.5 <i>Data collection</i> | 125 |
| 5.2.6 <i>Data analysis</i> | 129 |
| 5.3 Results | 134 |
| 5.3.1 <i>Respondent-rated difficulty</i> | 135 |
| 5.3.2 <i>Preference stability</i> | 137 |
| 5.3.3 <i>Dominant preferences and non-trading behaviour</i> | 138 |
| 5.3.4 <i>Choice analysis</i> | 139 |
| 5.3.5 <i>CSPC budget allocations</i> | 140 |
| 5.4 Identifying a preferred elicitation format | 142 |
| Appendix 5.1: Sample DCE and CSPC choice tasks | 146 |
| Appendix 5.2: DCE & CSPC choice model coefficients, marginal rates of substitution and importance ranks | 149 |
| Chapter 6: Primary data collection methods and sample characteristics | 151 |
| 6.1 Survey methods | 152 |
| 6.1.1 <i>Experimental design</i> | 153 |
| 6.1.2 <i>Data collection</i> | 155 |
| 6.1.3 <i>Choice and ratings tasks</i> | 157 |
| 6.2 Primary experimental design | 160 |
| 6.3 Sample characteristics | 161 |
| 6.3.1 <i>Responses by design block and version</i> | 165 |
| 6.3.2 <i>Completion times</i> | 166 |
| 6.3.3 <i>Respondent attitudes toward rationing</i> | 167 |
| 6.4 Implications for democratic or Communitarian priority setting | 170 |
| Appendix 6.1: Pilot preference weights used in developing the primary experimental design | 172 |
| Appendix 6.2: Primary experimental design, by block and version | 173 |
| Appendix 6.3: Sample choice tasks | 176 |
| Appendix 6.4: Rating tasks | 178 |
| Chapter 7: Comparison of the DCE and CSPC formats | 179 |
| 7.1 DCE-CSPC comparisons | 180 |
| 7.1.1 <i>Completion rates</i> | 180 |
| 7.1.2 <i>Respondent-rated difficulty and confidence</i> | 181 |

| | |
|--|------------|
| 7.1.3 Tests of non-satiation and stability | 182 |
| 7.1.4 Learning and fatigue effects | 183 |
| 7.1.5 Dominant preferences | 183 |
| 7.1.6 QALY maximisation | 185 |
| 7.2 DCE and CSPC response behaviours | 188 |
| 7.2.1 Questionnaire completion rates | 188 |
| 7.2.2 Respondent-rated difficulty and confidence | 189 |
| 7.2.3 Non-satiation and preference stability | 191 |
| 7.2.4 Learning and fatigue effects | 193 |
| 7.2.5 Dominant preferences | 196 |
| 7.2.6 QALY maximisation | 198 |
| 7.3 Discussion of the DCE-CSPC comparisons | 203 |
| Appendix 7.1: Post hoc test of significant ANOVA results | 207 |
| Appendix 7.2: Probit analysis of QALY maximising choices by task sequence and questionnaire format | 208 |
| Appendix 7.3: Distribution of choice-attribute correlations | 209 |
| Chapter 8: Primary DCE results | 211 |
| 8.1 Specifying the DCE model | 212 |
| 8.1.1 Agent vs. public preferences | 216 |
| 8.2 Modelling individual heterogeneity | 217 |
| 8.2.1 Relating individual characteristics to latent class membership | 222 |
| 8.3 Estimating welfare effects using compensating variation | 223 |
| 8.3.1 Scenario rankings | 229 |
| 8.4 DCE results | 230 |
| 8.4.1 Overall DCE results | 231 |
| 8.4.2 DCE scenario rankings | 234 |
| 8.4.3 DCE results by latent class | 239 |
| 8.4.4 Public vs. agent preferences | 244 |
| 8.5 Discussion of DCE results | 246 |
| Appendix 8.1: Alternative DCE models and value functions, by improving information criteria | 250 |
| Appendix 8.2: Combined latent class model coefficients | 252 |
| Appendix 8.3: Latent class model coefficients, by class | 253 |
| Appendix 8.4: Alternative DCE public-agent interaction value functions, by improving information criteria | 254 |
| Appendix 8.5: Dummy-coded MNL with agent interactions coefficients | 256 |
| Chapter 9: Primary CSPC results | 257 |
| 9.1 Specifying the CSPC model | 258 |

| | |
|--|------------|
| 9.2 Estimating welfare effects | 262 |
| 9.2.1 <i>Public vs. agent preferences</i> | 262 |
| 9.3 Scenario rankings | 263 |
| 9.4 Distributional preferences | 263 |
| 9.5 Comparison of DCE and CSPC welfare estimates | 264 |
| 9.6 CSPC Results | 265 |
| 9.6.1 <i>Compensating variations</i> | 268 |
| 9.6.2 <i>Marginal effects</i> | 269 |
| 9.6.3 <i>Public vs. agent preferences</i> | 270 |
| 9.6.4 <i>Scenario rankings</i> | 271 |
| 9.6.5 <i>Distributional preferences</i> | 275 |
| 9.6.6 <i>Comparison of preferences from the DCE and CSPC</i> | 278 |
| 9.7 Discussion of CSPC results | 281 |
| Appendix 9.1: Alternative CSPC models and utility functions, by improving information criteria | 285 |
| Appendix 9.2: CSPC double-bounded tobit coefficients | 286 |
| Appendix 9.3: Alternative CSPC public-agent interaction value functions, by improving information criteria | 287 |
| Appendix 9.4: CSPC double-bounded tobit with agent interactions | 288 |
| Appendix 9.5: Latent class multinomial logistic coefficients and differences by questionnaire format | 289 |
| Chapter 10: Discussion and concluding remarks | 291 |
| 10.1 An equity-efficiency trade-off? | 295 |
| 10.2 Comparison with other societal preference elicitation | 297 |
| 10.3 Choosing between the DCE and CSPC | 300 |
| 10.4 Implications for healthcare policy | 302 |
| 10.5 Methodological challenges to incorporating societal preferences into healthcare priority setting and suggestions for future research | 305 |
| 10.5.1 <i>Aggregating heterogeneous societal preferences</i> | 305 |
| 10.5.2 <i>Eliciting reliable preferences</i> | 308 |
| 10.5.3 <i>Public involvement in priority setting</i> | 309 |
| 10.6 Concluding remarks | 311 |
| Appendix 10.1: Summary of recent DCE and CSPC stated preference elicitation | 313 |
| References | 315 |

List of tables

| | |
|---|-----|
| Table 3.1: Potentially relevant attributes | 39 |
| Table 3.2: Attributes in recent stated preference elicitations | 66 |
| Table 3.3: Summary of the empirical ethics review | 69 |
| Table 4.1: Summary of stated preference elicitation methods | 100 |
| Table 4.2: Summary of recent DCE and CSPC methods | 105 |
| Table 5.1: Pilot survey attributes and levels | 115 |
| Table 5.2: Respondent characteristics by questionnaire | 134 |
| Table 5.3: Respondents rating the questionnaires ‘somewhat difficult’ or ‘extremely difficult’ to understand | 136 |
| Table 5.4: Respondents rating the questionnaires ‘somewhat difficult’ or ‘extremely difficult’ to answer | 136 |
| Table 6.1: Primary survey attributes and levels | 153 |
| Table 6.2: Experimental design attribute correlations | 161 |
| Table 6.3: Canadian and survey age-sex distributions | 162 |
| Table 6.4: Age and sex distribution by questionnaire design | 164 |
| Table 6.5: Unique respondents by design block and version | 165 |
| Table 6.6: Unique choices by design block and version | 166 |
| Table 6.7: Rationing attitudes by sample | 167 |
| Table 6.8: Proportion of public and agent respondents supporting stakeholder roles in healthcare funding decisions | 168 |
| Table 6.9: Proportions by comfort with having their preferences used in priority setting decisions | 169 |
| Table 6.10: General public support for a public role in decision making by own preference comfort | 170 |
| Table 7.1: Completion rates by questionnaire format and respondent subgroup | 188 |
| Table 7.2: Respondents rating the tasks as somewhat or extremely difficult to understand, by format | 189 |
| Table 7.3: Respondents rating the tasks as somewhat or extremely difficult to answer, by format | 190 |
| Table 7.4: Respondents who indicated they were somewhat or extremely confident that their answers accurately reflected their preferences, by format | 190 |
| Table 7.5: Non-satiation by questionnaire format and stakeholder group | 191 |
| Table 7.6: Preference stability by questionnaire format and respondent subgroup | 192 |
| Table 7.7: ANOVA adjusted p-values by choice set, format and block | 194 |

| | |
|--|-----|
| Table 7.8: Individuals with perfect choice-attribute correlation by attribute and format | 197 |
| Table 7.9: Individuals with confirmed dominant preferences by attribute and format | 198 |
| Table 7.10: Respondents by number of QALY maximising choices and questionnaire format | 199 |
| Table 7.11: Respondents by number of QALY maximising choices and agent status | 200 |
| Table 7.12: Respondents by number of individual and aggregate QALY maximising choices | 201 |
| Table 7.13: Predicted QALY maximising probabilities by task sequence and format | 202 |
| Table 7.14: Summary of DCE-CSPC comparisons | 203 |
| Table 7.15: Tukey's test of honest significant difference | 207 |
| Table 7.16: Probit model of likelihood of choosing the QALY maximising alternative | 208 |
| Table 8.1: Initial and final health state interaction values | 213 |
| Table 8.2: Wald tests of non-linearity in dummy-coded parameters | 232 |
| Table 8.3: DCE compensating variations by change in attribute levels | 232 |
| Table 8.4: DCE scenario rankings by predicted utility and probability of choice | 235 |
| Table 8.5: Age-stratified DCE scenario rankings by predicted utility and probability of choice | 237 |
| Table 8.6: Wald tests of non-linearity in dummy-coded parameters, by latent class | 241 |
| Table 8.7: Compensating variations and differences between latent classes by attribute change | 242 |
| Table 8.8: Compensating variations and differences for agents and the general population | 245 |
| Table 9.1: Initial and final health state differences interaction values | 261 |
| Table 9.2: CSPC compensating variations by attribute differences | 268 |
| Table 9.3: CSPC double-bounded tobit marginal effects | 270 |
| Table 9.4: CSPC scenario rankings by predicted difference in utility | 272 |
| Table 9.5: CSPC scenario rankings controlling for relative age | 274 |
| Table 9.6: Respondents by the number of tasks equalising or maximising the budget allocation | 276 |
| Table 9.7: Respondents by the number of tasks equalising patients treated or QALYs gained | 277 |
| Table 9.8: Compensating variations and differences by attribute and questionnaire format | 280 |

List of figures

| | |
|--|-----|
| Figure 5.1: Attribute relative importance by format | 140 |
| Figure 5.2: Pilot CSPC budget allocations | 141 |
| Figure 6.1: DCE and CSPC completion times, general population respondents only | 166 |
| Figure 7.1: Distribution of individual budget differences between the repeated CSPC tasks | 193 |
| Figure 7.2: Choices and budget allocations by design, choice set and task sequence | 194 |
| Figure 7.3: QALY maximising choices by questionnaire format | 199 |
| Figure 7.4: Predicted probabilities of choosing/prioritising the QALY maximising alternative by task and questionnaire format | 208 |
| Figure 8.1: Illustrating compensating variation | 224 |
| Figure 8.2: Compensating variation with quasilinear utility | 226 |
| Figure 8.3: DCE compensating variations by attribute | 233 |
| Figure 8.4: Latent class 1 membership probability density | 239 |
| Figure 8.5: DCE compensating variation by attribute and latent class | 243 |
| Figure 8.6: Compensating variations for changes in initial and final health state, by group | 246 |
| Figure 9.1: Primary CSPC budget allocations | 265 |
| Figure 9.2: Pooled tobit residuals | 266 |
| Figure 9.3: CSPC compensating variations by attribute | 269 |
| Figure 9.4: CV differences by attribute change, DCE vs. CSPC | 279 |

Acronyms

| | |
|-------|--|
| BWS | Best-worst scaling (conjoint ranking) |
| CADTH | Canadian Agency for Drugs and Technology in Health |
| CSPC | Constant-sum paired comparison |
| CSS | Constant-sum scaling |
| CV | Compensating variation |
| DCE | Discrete choice experiment |
| ME | Magnitude estimation |
| MNL | Multinomial logit |
| MRS | Marginal rate of substitution |
| NICE | National Institute for Health and Care Excellence (UK) |
| PTO | Person trade-off |
| SP | Stated preferences |
| RP | Revealed preferences |
| RUT | Random utility theory |
| QALY | Quality-adjusted life year |

Chapter 1: Introduction, objectives and thesis outline

Health is a primary foundation of what Culyer (2001b) has termed “a flourishing life,” the ultimate human condition. Any reduction in health, through disease or injury, reduces a person’s potential to enjoy such a life. Others have been more specific, defining health as part of a set of capabilities that provide an individual a normal range of opportunity (Sen 1985; Daniels 2001). This fundamental importance imbues health with a special moral significance to society (Sen 2002; Anand 2002), and in turn gives healthcare a particular significance, as it is an important – although not the only – factor in achieving and maintaining an optimal level of health (Culyer 2001b).

Modern healthcare is able to offer some health benefit to almost any condition, but this very effectiveness suggests that the demand for healthcare is likely to outstrip supply (Appleby & Harrison 2006; New 2000). In a market system, demand would be constrained by the price mechanism and an individual’s willingness and ability to pay. However, there are a number of specific and well recognised failures in the market for healthcare. These include uncertainty around the timing and quantity of an individual’s demand for healthcare, the ‘public good’ nature of many healthcare services, positive externalities associated with healthcare, asymmetry of information between patients and providers, and the absence or distortion of price signals (Arrow 1963). While these market failures are not necessarily unique to the health sector, most believe that healthcare is so fundamentally different than other goods and services that a market mechanism would fail to deliver an efficient or

equitable level of health (Daniels 2001; Culyer 2001a; Maynard & Bloor 1998; Hauck et al. 2004).

In light of these market failures, most governments in the world have undertaken to provide, to a greater or lesser degree, publicly funded healthcare to their citizens. However, even government resources are finite, so there must still be a mechanism for coping with excess demand. If societal healthcare resources are not to be allocated on the basis of the price mechanism, a process of rationing or priority setting¹ is required, which can be understood as “the deliberate and systematic withholding of beneficial goods or services on the grounds that society cannot afford to extend them.” (Fleck 1992) Through this process, effective healthcare must be denied to someone that could potentially benefit, and thus the fundamental problem facing the healthcare decision maker is how to decide who will be allowed to benefit from societal healthcare resources and who will not (New 1996). This thesis describes a normative economics approach to addressing this decision problem.

1.1 Thesis objectives

The thesis takes the position that an explicit approach to healthcare priority setting, based on clearly defined objectives and criteria that reflect the preferences of society, can improve the value that society derives from healthcare.² Value in this context should be understood as a broader concept than health, as the total value that society derives from healthcare may be greater or less than the sum of the value that individual patients derive from their own health gains. The degree to which these concepts differ reflects the societal desire for equity or distributive justice in the allocation of health gains, as for

¹ Although these terms are effectively equivalent and each may appear throughout the thesis, ‘priority setting’ will be preferred as it is more consistent with Broome’s (1989) view, adopted here, that fairness requires that resources should be allocated according to the strength of one person’s claim *relative* to another’s. In this view, it is not a question of which patient is treated and which is not, but of which patient gets priority.

² The terms healthcare and health gain will be used more or less synonymously when referring to the source of societal value, on the presumption that the primary output, and object of value, of healthcare is health gain. This is not true in the presence of caring externalities, where healthcare may also be valued for non-health outcomes such as dignity, compassion or maintenance of hope. These externalities are assumed away for now, but will be discussed later in the thesis.

equity reasons society may value health gains to some patients more (or less) highly than gains to others. This implies that society may be willing to sacrifice some degree of efficiency in maximising aggregate health gains in exchange for a distribution that is perceived to be more fair; this is known as the ‘equity-efficiency trade-off’ (Wagstaff 1991; Sassi et al. 2001). The more strongly society prefers a particular conception of equity, the greater the sacrifice in terms of potential health gains it should be willing to make to achieve that distribution.

The key challenge in this approach to maximising the value of healthcare is defining the criteria by which value should be judged. Within healthcare, value has conventionally been defined by decision makers in terms of quality-adjusted life years (QALYs), which weight years of life by a quality adjustment representing the ‘healthiness’ of those years (Culyer 1989; Brouwer et al. 2008; Coast 2009). Under a QALY maximising objective and a presumption of distributive neutrality, where the societal value of an additional QALY is held to be the same regardless of who receives it (Nord et al. 1995; Dolan et al. 2005), priority has been given to patients with conditions whose treatment will generate the greatest QALY gains.

Although this QALY maximising decision rule may be entirely consistent with societal preferences, it will be argued that this relatively narrow definition of value appears to neglect a number of patient and program³ characteristics that empirical studies of societal preferences have suggested may be relevant. Explicit consideration of these factors, through a broader conception of value, could align the allocation of resources more closely with societal preferences. This, in turn, would lead greater efficiency in translating healthcare resources into societal value, as well as a more equitable distribution of societal resources. It may also lead to greater trust and legitimacy in the priority setting process. To this end, **the primary objective of the thesis was to identify the factors relevant to the societal value of healthcare, and to estimate the strength of the equity-efficiency trade-off over these factors.** Secondary objectives were to compare different methods for eliciting these societal preferences, and to test the

³ As the Canadian usage of ‘program’ rather than the British ‘programme’ was presented to survey respondents, this form will be used throughout the thesis.

homogeneity of preferences between the public and the decision-making agents responsible for making priority setting decisions on their behalf.

1.2 Thesis outline

The thesis is divided into two parts. **Part one** provides a background on priority setting in healthcare, including an introduction to priority setting approaches within normative economics, a review of patient and program characteristics that may be relevant to priority setting, and a comparative review of the different stated preference methods that can be used to elicit the strength of societal preferences for different characteristics, particularly in the context of healthcare. **Part two** presents empirical work. This includes a pilot survey, which compared two different preference elicitation with the intention of identifying a preferred method, and a primary survey, which elicited the strength of societal preferences for the factors identified in part one from an age and gender representative sample of the Canadian public as well as a convenience sample of self-identified decision-making agents. The results of the pilot survey did not indicate a clearly preferred elicitation method, so the same two stated preference methods tested in the pilot survey were used in the primary survey. Part two therefore also includes a comparison of the response behaviours of two methods based on the larger sample of the primary survey, and a detailed discussion of the relative preferences derived from each method. The thesis concludes with a discussion of the results and their implications for healthcare policy, as well as the limitations of this work and suggestions for future research. A more detailed outline of the chapters in each section is presented below.

It is important to highlight that whereas respondents to pilot survey were told that the health states in the survey were entirely hypothetical, respondents to the primary survey were told that the different patient groups all had some form of cancer. A cancer context was used for pragmatic reasons, as funding for the primary survey was provided by the Canadian Centre for Applied Research in Cancer Control, but a specific context may also provide respondents with a more concrete and more comparable understanding of the different health states presented in the survey. Indeed, the impact of cancer and cancer treatments on

morbidity and mortality will be reasonably familiar and understandable to most respondents. However, to ensure a focus on the attributes and levels of each program and not the disease labels, the alternatives in each choice task were unlabelled and presented generically as Program A and Program B. In the absence of specific labels, there is little reason to suspect that the results from the primary elicitation should not be generalizable to other disease contexts.

1.2.1 Part one

Chapter 2 offers an introduction to normative economics, and contrasts welfarist and extra-welfarist approaches to normative economic decision making. The welfarist approach emphasises individual well-being, and as such, it is argued that it offers an impractical guide to the allocation of societal resources. The extra-welfarist approach, in theory, goes beyond individual well-being and allows for a broader understanding of societal well-being. However, this requires an implicit or explicit definition of the factors that may contribute to societal well-being, as well as an understanding of who should contribute to that definition. The chapter will outline the arguments for and against an explicit definition of these factors, and discuss the merits of narrow impartiality and objectivity versus broader and more subjective perspectives in societal priority setting. The chapter will also discuss the use of the equity-weighted QALY as one approach to explicitly incorporating societal preferences into healthcare priority setting.

Chapter 3 reviews the potential factors that may contribute to the societal value of healthcare. The most straightforward approach to identifying these factors is to ask people which factors they consider important. However, many argue that simple majority support for particular attributes or characteristics is not sufficient grounds for distributing something as important as healthcare. Therefore, the review took an empirical ethics approach, “involving both an empirical study of population values and ethical analysis of the results,” (Richardson & McKie 2005) to identify factors that can be considered both *relevant* and *fair*. To this end, attributes had to have empirical evidence of public support, and be consistent with a dominant theory of distributive justice.

Chapter 4 reviews different stated preference methods for eliciting the strength of societal preferences. The empirical ethics review of Chapter 3 was not sufficient to justify priority for particular factors, as most empirical studies gave little or no consideration to the trade-offs between factors or outcomes. Rather, estimating the relative strength of preferences requires a process that forces trade-offs between these factors. This chapter compares different stated preference methods, and concludes that two methods – discrete choice experiments (DCEs) and constant-sum paired comparisons (CSPCs) – appear to have advantages in eliciting societal preferences in this context.

1.2.2 Part two

Chapter 5 details a pilot survey used to compare the DCE and CSPC elicitation methods to identify a preferred stated preference method for the primary elicitation. The survey was also used to refine the wording and presentation of the choice tasks. The chapter outlines the methods used in developing the survey, including the assignment of levels to the attributes identified in Chapter 3, the development of the experimental design, and the data collection and analysis. The two stated preference methods were compared on a number of dimensions, and the results of these comparisons, particularly with respect to their bearing on identifying a preferred elicitation method for the primary survey, are also detailed.

Chapter 6 describes the methods used for the primary survey. The pilot survey identified advantages with both elicitation methods, and it was decided that it would be of interest to compare them in more detail based on the larger sample of the primary survey. As a result, both the DCE and the CSPC elicitation formats were used in the primary survey. This chapter emphasises the methodological differences from the pilot survey, including the survey sample, the experimental design, and the presentation and context of the choice tasks. As the following chapters present the results of the two elicitation formats separately, Chapter 6 also takes the opportunity to present a summary of the overall survey sample, including their representativeness of the larger Canadian population and their attitudes towards rationing and their support for public involvement in priority setting. The chapter concludes with a discussion of the

implications of their attitudes for more participatory approaches to priority setting.

Chapter 7 presents a comparison of the two stated preference methods, based on the larger, representative sample of the primary survey. The implications of these comparisons for a preferred method for eliciting societal preferences are also discussed.

Chapters 8 and 9 present the methods used in the statistical modelling of the DCE and CSPC choice responses and estimating the welfare effects associated with changes in the attributes included in the elicitations. The results, in terms of marginal welfare effects and holistic scenario rankings are presented, along with a comparison of the preferences of the general public and decision making agents.

Finally, **Chapter 10** discusses the implications of these results for the allocation of societal healthcare resources, and for the use of QALY maximisation as a societal decision rule. It also compares these results with previous elicitations, discusses the strengths and limitations of the methods and results, and outlines how future research may be able to build upon the strength and address the limitations.

Chapter 2: Normative economics and healthcare priority setting

Normative economics addresses the question of how resources *ought* to be distributed, weighing the maximisation of outcomes against the ‘fairness’ of the distribution, based largely on ethical and philosophical visions of distributive justice (Culyer 2001a; Johansson-Stenman 1998). Unlike positive economics, which is in principle a value-free description of *what is*, normative economics, by definition, starts with an implicit or explicit value judgement about what is ‘good’ or ‘desirable’ to describe *what ought to be* (Feldman & Serrano 2006; Johansson-Stenman 1998).

This chapter describes the two dominant approaches to normative economics: the welfarist approach, described in section 2.1, and the extra-welfarist approach, described in section 2.2. The welfarist approach emphasises individual well-being, while the extra-welfarist take a broader view and emphasises societal well-being. However, this requires some definition of the factors beyond individual well-being that contribute to societal well-being, as well as an understanding of who should contribute to that definition. Section 2.3 outlines the arguments for and against an explicit definition of these factors, while section 2.4 describes more and less inclusive approaches to defining which potential factors may be relevant to the societal value of healthcare, and discusses the role of objectivity in societal priority setting. Finally, section 2.5 describes the equity-weighted QALY as one approach to explicitly incorporating societal preferences in healthcare priority setting.

2.1 The welfarist approach

Hurley (1998) and Brouwer et al. (2008) describe four value judgements that make up the neo-classical welfarist approach to normative economics: utility maximisation, individual sovereignty, consequentialism and welfarism. The principle of utility maximisation implies that individuals maximise their welfare by comparing different alternatives and choosing the one with the greatest 'utility,' which should be understood as an ordinal measure of the degree to which a particular alternative satisfies an individual's preferences. A more preferred alternative is said to have greater utility, and consistently choosing alternatives with the highest utility is assumed to maximise an individual's overall welfare. Individual sovereignty holds that welfare (or utility) is unique to an individual, and that the individual can be the only judge of their own welfare. This principle rejects paternalism, or that a third party may know better than the individual what is best for them. Consequentialism holds that any action or decision must be judged solely by its outcome, not the processes or intentions that led to that outcome. Finally, welfarism holds that the 'goodness' of any situation should be judged solely by the utility attained by individuals in that situation. The primacy of individual preferences in neo-classical economic theory is based on the assumption that individuals are rational, self-interested and perfectly informed; thus, individuals will prefer X to Y if, and only if, X is in fact better for them. This leads to a formal theory of welfare that holds that the welfare of an individual can be equated with how well their preferences are satisfied (Feldman & Serrano 2006; Hausman & McPherson 2009).

The welfarist approach shares the principles of utility maximisation, consequentialism and welfarism with utilitarianism. But whereas utilitarianism takes the view that "justice is ultimately a matter of maximising the sum total of human happiness" (Mill 1871; Williams & Cookson 2000), and that alternatives should be evaluated on the basis of aggregate individual utility, the welfarist approach is adamant that utility is ordinal, and cannot be compared or aggregated across individuals (Brouwer et al. 2008). Individual sovereignty and welfarism effectively rule out interpersonal comparisons – individuals are to be the sole judges of their welfare and the welfare of each individual is equally important. Within the welfarist framework, therefore, the societal desirability of

a reallocation is judged by the Pareto Improvement Criterion, which states that a potential reallocation is a desirable improvement if, *and only if*, the welfare of at least one member of society is improved without making anyone worse off (Sugden & Williams 1978; Feldman & Serrano 2006). The current allocation of resources is taken as a given, and if resources cannot be reallocated in a way that satisfies this criterion, the current allocation is said to be ‘Pareto optimal.’ As this may rule out reallocations that could improve aggregate societal welfare, the welfarist approach has been described as applying a weak version of utilitarianism, in that it is willing to accept as optimal an allocation that does not maximise aggregate welfare (Culyer 2001a). The strict reallocation conditions of the Pareto Improvement Criterion also mean that the welfarist approach cannot accommodate equity concerns – the well-being of the worst-off in society can be no more (or less) important than the well-being of the best-off. This has the implication that flagrantly unequal or inequitable allocations can be considered Pareto optimal if the existing distribution of resources cannot be reallocated without creating a ‘loser’ (Hurley 1998; Konow 2003; Feldman & Serrano 2006).

2.2 The extra-welfarist approach

Although the Pareto criterion is in itself a relatively weak and uncontroversial value judgement, the supremacy of the individual means that it is a restricted and somewhat impractical guide to allocating societal resources, which generally involves reallocating resources from the better-off to the worse-off (Hauck et al. 2004; Feldman & Serrano 2006; Tsuchiya & Williams 2001; Coast et al. 2008b). As a result, many of the principles of the welfarist framework have been modified to provide more practical normative guidance to societal decision making. This has led to ‘extra-welfarist’ or ‘non-welfarist’ approaches (Brouwer et al. 2008; Hurley 1998; Culyer 1989; Coast et al. 2008b).⁴

⁴ The distinction between the terms extra-welfarist and non-welfarist is not always clear, and the two are often used more or less synonymously, but Coast (2009) offers a useful perspective in suggesting that extra-welfarism can be seen as a specific theoretical framework within the larger set of often atheoretical non-welfarist approaches.

There are four key principles that distinguish the extra-welfarist approach from the welfarist approach (Brouwer et al. 2008; Hurley 1998; Culyer 1989). First, it allows for the consideration of non-utility factors as well as individual utility. Second, it incorporates valuations from sources other than the affected individual. This allows for external value judgements that may override the principle of individual sovereignty. Third, it allows for the explicit incorporation of equity weights that are not necessarily preference based. Fourth, it assumes that utility is cardinal, and allows for inter-personal comparisons of well-being. The extra-welfarist approach moves toward a concept of societal well-being that Culyer (1989) argues “transcends traditional welfare” by supplementing information on individual welfare with information on other aspects of individuals, including the distribution of well-being between them. Hurley (1998) goes further, and suggests that non-utility factors may even be more important than individual utility. These principles – and particularly the inter-personal comparison of cardinal utilities – allow for a relaxed version of the Pareto improvement criterion, known as the potential Pareto improvement, or Kaldor-Hicks criterion. If, *in principle*, the gainers from a particular reallocation are able to fully compensate the losers and remain at least as well off as before the reallocation, the new state is considered a potential Pareto improvement over the original state (Feldman & Serrano 2006; Tsuchiya & Williams 2001).

The potential for a redistribution that would leave everyone at least as well off is used as a justification within both the welfarist and extra-welfarist approaches for emphasising the maximisation of outputs and disregarding the distribution of those outputs as a political matter (Sugden & Williams 1978; Coast 2009). However, Sassi et al. (2001) argue that in neglecting equity concerns, economics loses much of its normative power and restricts itself to the relatively narrow domain of technical efficiency. Furthermore, within a healthcare context, it is not possible to separate the production of health from its distribution; production and allocation happen simultaneously (Coast 2009). As Menzel (1999) points out, “...it is often not possible to redistribute health, or to compensate for healthcare allocations through the distribution of other goods. It is difficult to compensate someone who has died because one program received priority over another.” For these reasons, healthcare priority setting can be seen

as a matter of allocative as well as technical efficiency, and as such requires explicit consideration of equity and distributive justice (Williams 1988; Menzel et al. 1999; Coast 2009).

In this context, Hurley (1998) describes the ‘analytic imperative’ of the extra-welfarist approach as follows: from the characteristics of people, define a set of normatively relevant characteristics; measure the relative level deprivation within those characteristics and the corresponding need⁵ for commodities (e.g. healthcare) to address these deprivations; and compare alternative allocations of commodities with respect to their ability to alleviate deprivations. This description of defining normatively relevant characteristics and comparing alternative allocations highlights (at least) two questions that must be resolved before proceeding with an extra-welfarist evaluation: on what terms should alternative allocations be compared, and who should define those terms? The first question concerns the explicitness of the decision rule for choosing between allocations, and the second question concerns the inclusiveness and perspective of the priority setting process. These two issues will be considered in turn below.

2.3 Explicitness in priority setting

Approaches to healthcare priority setting can be understood as implicit or explicit. Coast (1997) defines an implicit approach as the rationing or prioritisation of healthcare where neither the decisions about what programs to fund nor the bases of these decisions are clearly expressed. Under an implicit approach, equity-efficiency trade-offs are implicitly recognised but not explicitly quantified, and prioritisation decisions are based largely on the judgement of individual decision makers. Under more explicit approaches, the responsibility of decision makers is to define a consistent and transparent set of factors and weights that define acceptable equity-efficiency trade-offs, and prioritisation decisions are made on the basis of these weights and a pre-defined decision rule rather than individual judgement.

⁵ Hurley (1998) noted that deprivation does not automatically imply a corresponding need for healthcare, as need also requires an effective treatment. In the absence of an effective treatment, a person cannot be said to have need.

2.3.1 Implicit priority setting

Proponents of a more implicit approach see priority setting as an “inescapably political process” (Ham & Coulter 2001), requiring discussion and compromise rather than inflexible decision rules (Hunter 2001; Robinson 1999). A key benefit of an implicit approach is the avoidance of conflict:

Principles that incorporate semiautomatic formula for implementing them (like maximising health benefits) tend to be highly contentious, while uncontentious principles owe their acceptability to the fact there is ambiguity about their implementation (Klein 1997).

Many argue that this ambiguity, in terms of what is funded and why, offers the flexibility necessary to address the inherent complexity of healthcare decision making, including the practical difficulties of defining and weighting explicit criteria, and enforcing the resulting decisions across all settings (Hunter 2001; Klein 1997; Mechanic 1995). A lack of transparency is also argued to be necessary to overcome consumer and provider resistance and lobbies (Klein 1992). In this view, a lack of transparency allows decision makers to make the ‘correct’ choice rather than the ‘popular’ choice. This is similar to Wirtz et al.’s (2003) suggestion of a “hidden curriculum” within healthcare decision making that tacitly emphasises process concerns over technical factors such as efficiency, effectiveness and affordability. In their view, process factors such as the maintenance of good relations with major stakeholders (what they refer to as ‘picking your battles’), the management of organizational burden (managing trust and morale, in addition to purely financial issues) and public defensibility (emphasising perceived fairness over technical measures) justify taking a more implicit approach to priority setting. Such an approach is consistent with a cost-consequence decision framework, where the costs and benefits are measured and presented in a disaggregated format, but each decision maker assigns his or her own weights across the different factors in deciding whether the benefits of a particular program are worth the costs (Mauskopf et al. 1998).

At the individual level, some proponents of an implicit approach also noted that it could be painful for patients to be told that effective care is being denied, and for decision makers to take responsibility for such decisions. From a utilitarian perspective, an implicit approach may maximise societal well-being by

minimising such ‘deprivation disutility’ and ‘denial disutility,’ respectively (Coast 1997; Mooney & Lange 1993). They suggested that the patient and the clinical decision maker can only be made worse-off by the explicit communication that effective and beneficial healthcare was denied on the basis of criteria other than clinical effectiveness. Therefore it is better for both parties, and for society in general, to leave the patient with the impression that the decision was based on clinical factors beyond anyone’s control (Coast 1997). Qualitative research, though, has found that even though patients acknowledged some distress from the knowledge, they consistently expressed a desire to be told if their care was being rationed (Coast 2001b; Owen-Smith et al. 2010). The primary motivation appeared to be a simple desire to be as informed as possible about their care, and to have “a good explanation as to why the decision was made.” (Coast 2001b) Patients as well as providers also felt that the knowledge an effective treatment was available but had been rationed would allow patients to seek the treatment by other means, such as political lobbying, or private or self-funding (Coast 2001b; Owen-Smith et al. 2010). From the perspective of the providers, although most expressed support for a principle of full and explicit disclosure, many acknowledged being less explicit about rationing decisions when they felt that a patient may not have had alternative means to access treatment (Coast 2001b; Owen-Smith et al. 2010).

2.3.2 Explicit priority setting

Proponents of a more explicit approach argue that from an ethical and moral perspective, clearly defined objectives and criteria, and transparency in the decision making process, are the bases of citizens’ democratic rights to informed consent and political autonomy. It is also the basis of citizens’ ability to hold decision makers responsible for their decisions (Doyal 1997; Lauridsen et al. 2007; Rumbold et al. 2012). As Doyal (1997) notes, transparency and accountability may undoubtedly lead citizens to give decision makers a difficult time, but that is their right in a democracy, particularly over issues with the fundamental importance of healthcare. Furthermore, as denial disutility is at least in part a consequence of decisions around levels of taxation and funding, it makes little sense to hide the necessity of priority setting from citizens, as they

can only make informed choices about funding levels if they can see the consequences of their decisions (Buxton & Chambers 2011). As an aside, it is interesting to note that one result of the Oregon experiment in explicit priority setting of the early 1990s was an increase in the overall level of healthcare funding (Ham 1998).

From a technical perspective, proponents of more explicit approaches argue that bringing as much relevant information as possible together within an explicit framework supports rigorous evaluation and continuous improvement to a much greater degree than implicit approaches (Dowie 1998; Doyal 1997; Mitton 2002). Doyal (1997) argues that to not make an attempt be explicit in decision criteria is to give up the ability to evaluate the efficiency or justice of a particular distribution of resources, and to accept the possibility that a redistribution could do as much harm as good. Finally, making the criteria for decisions more transparent decreases the potential influence of special interest groups and may increase trust in the decision-making process (Coast 2001b; Devlin et al. 2003; Doyal 1997). Fleck (1992) suggests that implicit priority setting can create an invisible class of ‘others,’ who may be victims of injustice without knowing it. Similarly, Broqvist and Garpenby (2014) suggest that priority setting is based on a social contract by which citizens accept the need to forego some effective healthcare in order that those with a greater need may receive priority; in return they expect that others will stand aside when they have the greater need. A poor understanding of why particular patients were prioritised erodes trust in this contract, and makes citizens less willing to stand aside for others.

A more explicit approach to priority setting appears to be associated with a more informed citizenry, more accountable decision makers, greater opportunities for evaluation and improvement, and greater trust in the priority setting process. To the extent that these outcomes are in themselves desirable, a more explicit approach to healthcare priority setting appears justified. However, it is still necessary to define what factors will be considered in an explicit decision-making approach, and perhaps even more importantly, who will define these factors.

2.4 Inclusiveness and objectivity within the extra-welfarist framework

Coast et al. (2008b) explain that because an individual's preferences are not paramount within the extra-welfarist approach, it is necessary to decide what other factors are normatively relevant and what weight each should carry in the decision making process. This is what Broome (1989) describes as distinguishing an individual's normative *claims* to some good or resource from the *reasons* they should have it. He argues that claims, and not reasons, are the object of fairness: "if there are reasons why a person should have a good, but she does not get it, no unfairness is done her unless she has a claim to it." Critically, it is also necessary to decide who should define what characteristics are relevant; that is, which reasons rise to the level of claims and which do not (Broome 1989). There is a range of perspectives that can be applied, but this range is arguably anchored at one end by the strictly impartial decision-maker perspective, and at the other by a more inclusive and subjective democratic or Communitarian perspective.

2.4.1 The decision-maker perspective and QALY maximisation

The extra-welfarist approach has most commonly adopted what Sugden and Williams (1978) call a 'decision maker' perspective, whereby the relevance of different characteristics is defined by those individual responsible for making (or analysing) policy decisions on behalf of society (Sugden & Williams 1978; Coast 2004). A perceived advantage of the decision maker perspective is that societal decision makers, on the basis of their knowledge, expertise and professionalism, are uniquely "impersonal, impartial, unbiased and neutral" (Buchanan et al. 1998), or in other words, objective. Indeed, when Coast et al. (2001a) asked members of the general public and a group of healthcare decision-makers, including government bureaucrats, physicians, hospital administrators, who should participate in healthcare rationing, they found that decision makers as well as the public felt that citizens lacked sufficient objectivity. Both groups viewed objectivity as the ability to make decisions based solely on facts while setting aside any emotion or empathy.

Relying on impartial decision makers to make societal decisions is an example of 'procedural objectivity,' or the idea that objective decision makers will tend to reach an objective truth. In this context, Fine (1998) defines an

objective truth as one that can be accepted by all concerned with no further persuasion or explanation. This is in contrast to a subjective truth, which may be true from the perspective of a particular individual, but not necessarily true for all individuals. An example of an objective truth is that 10 is a larger number than 9; an example of a subjective truth is that blue is a better colour than red. Like blue versus red, the optimal allocation of healthcare resources is not an objective truth, and the value of a particular allocation ultimately rests upon subjective tastes, perspective and persuasion (Fleck 1992; Klein & Williams 2000; Daniels 2001). Relying on small groups of professional decision makers is viewed as a way to resolve this dilemma, and to arrive at an allocation that is objectively ‘best.’ Although this approach concentrates decision making authority in the hands of a relatively small group of decision makers, Sugden and Williams (1978) suggest that such an approach is fair and representative of broader society to the extent that these decision makers occupy their position as a result of a socially accepted process, and to the extent that they can be held accountable for their decisions through the same process. Brouwer, Culyer, van Exel and Rutten (2008) go even further, and suggest that the responsibility of societal decision makers is not to reflect how citizens *would* act, but rather how they *ought* to act, avoiding what Robinson (1999) refers to as a “dictatorship of the uninformed.” This is consistent with the view that less transparency allows decision-makers to the correct choice rather than the popular choice.

Within this decision maker perspective, aggregate health rather than individual utility has tended to be paramount. Coast et al. (2008b) argue that this perspective has been strongly influenced by Sen’s Capability theory, which holds that an individual’s well-being should be judged not by their own subjective utility, but by their objective capability to do things that he or she has reason to value (Sen 2011). Sen (1992) argues that the welfarist conception of utility suffers in particular from problems of physical condition neglect and valuation neglect. Physical condition neglect suggests that a disabled person may adjust their expectations downward to accommodate their circumstances – what Sen describes as learning to take pleasure in small mercies. Although such an individual may have a high subjective utility relative to their lowered expectations, what should matter in evaluating societal utility is the individual’s

objectively limited range of capabilities (Mooney 2005; Richardson & McKie 2005). Similarly, valuation neglect implies that “the strength of desire is influenced by considerations of realism in one’s circumstances,” and therefore welfarism has an over-reliance on “what people ‘manage to desire’” and is “particularly neglectful of the claims of those who are too subdued or broken to have the courage to desire much.” (Sen 1992; Mooney 2005) So while utility in the welfarist approach is defined by an individual’s *subjective* reaction to their choices and desires, Sen’s conception of well-being is defined by the *objective* range of choices and desires available to an individual, avoiding a reliance on the ‘metric of desire.’ (Cookson 2005; Brouwer et al. 2008) By insisting that we must not value only happiness, Sen justifies a definition of well-being largely external to the preferences and desires of the individual (Sen 1992; Cookson 2005; Coast et al. 2008b). Indeed, Sugden (1993) suggests that Sen wants to say that some functionings are intrinsically valuable, whether they are desired or not.

Sen (1992; 2011) has resisted an explicit definition of what capabilities should be valued, saying that the relevant capability set will depend on the nature of the question being addressed, but Nussbaum (2011) has suggested that life and health are fundamental, and Culyer (1989) makes specific reference to Capability theory in discussing the development of the extra-welfarist approach to health economic evaluation. Culyer notes a broad range of potentially relevant characteristics, including a person's genetic endowment of health, relative deprivation, moral ‘worth’, pain, stigmatisation and relationships, but acknowledges that “the extra-welfarist approach has taken ‘health’ as the proximate maximand,” where health is most often measured in terms of the QALY. This approach has become known as QALY maximisation.

By equating well-being with health, and health with the QALY, it follows that the QALY is a merit good, with an intrinsic value outside of its contribution to an individual’s utility (Culyer 2001b; Dolan 2001; Gold 1996). By focusing on QALYs rather than individual utility, and – critically – by presuming that an additional QALY is of equal value to everyone (Nord et al. 1995; Dolan et al. 2005; Weinstein et al. 2009), QALY maximisation avoids the welfarist implication that resources should be directed away from those who may place a lower value on their health (Wagstaff 1991). QALY maximisation also

presumes 'distributive neutrality,' or that the value society derives from each additional QALY is the same regardless of the characteristics of who receives it or the number of QALYs they may have already gained (Nord et al. 1995; Dolan et al. 2005). However, this conflation of QALYs and well-being, along with the Potential Pareto Improvement Criterion's emphasis on the maximisation over the distribution of gains, imposes a narrow perspective where more QALYs is always necessarily better than fewer, and rules out trading health for other goals such as gains in individual utility or distributive justice, even if such a trade would increase overall well-being (Gold 1996; Dolan 2001; Coast 2009).

Despite the consistency between QALY maximisation and procedural objectivity, in the sense that the QALY was defined by impartial analysts as an objective measure of (health-related) well-being, this has not lead to its widespread acceptance as a societal decision rule (Drummond et al. 2003; Hoffmann et al. 2002; Innvaer et al. 2002; Ross 1995). This perhaps relates to Fine's (1998) characterisation of procedural objectivity as "the view from nowhere, and of no-one in particular." By carefully excluding personal perspectives from societal allocation decisions, he argues that procedural objectivity makes it impossible to understand the very nature of subjective truths: that truth depends on tastes, perspective, and persuasion. In his view, a societally preferred allocation of resources cannot be reached by means of procedural objectivity alone, and personal perspectives – particularly concerning visions of distributive justice – must be acknowledged.

He goes on to argue that the fundamental point of objectivity in societal decision-making is not *truth*, but *trust*. Citizens do not value objectivity because they believe it arrives at an objective truth; they value it because they believe it arrives at a decision they can trust. In this view, objectivity represents anything that improves trust in a decision. In some circumstances, trust may be enhanced by the impartiality of societal decision makers, but in others, trust may be enhanced by a broader process, with more personal perspectives. Similarly, Sen (2011) wonders if it is possible to have a "...satisfactory understanding of ethics in general and justice in particular that confines its attention to some people and not to others, presuming – if only implicitly – that some people are relevant while others simply are not?" In his view, 'universality of inclusion' is an

integral part of objectivity. Heldke and Kellert (1995) make a similar argument, and suggest that ‘pure’ objectivity – which they define as knowledge that is independent from the perspectives of particular persons – is impossible, and to a large extent, undesirable. They argue that “knowledge is actually strengthened by systematically increasing the number of concrete, identifiable perspectives represented.” Together, these views are consistent with a more directly democratic (Fleck 1992) or Communitarian (Callahan 2003a; Mooney 2005) approach to defining the normative relevance of different characteristics.

2.4.2 A democratic or Communitarian perspective

Fleck (1992) argues that in order to justify its rationing decisions, a democratic government must ultimately appeal to some vision of distributive justice. However, as a single view of justice is unlikely to be endorsed by all citizens – particularly those who may lose out as the result of a rationing decision – it is essential that the government be able to demonstrate the moral legitimacy of its particular vision. *Prima facie* moral legitimacy could be achieved, he suggests, by creating “social processes through which rationing decisions become something that we collectively impose upon ourselves.” To this end, he argues that it should be the responsibility of all citizens in a democracy to contribute to determining the fair allocation of scarce healthcare resources. Fleck acknowledges the difficult and uncomfortable choices that this process may require of citizens, but emphasises the responsibilities, as well as the rights, of citizens in a democracy.

The democratic approach outlined by Fleck is similar to the Communitarian approach, advocated by Mooney (1998b; 2005) and Callahan (2003a; 2003b). Callahan (2003a) rejects the individualistic principles of welfarism on the grounds that they preclude a *societal* understanding of well-being. Such principles, he argues, only make sense if one believes in an ‘invisible hand’ that can shape individual well-being into societal well-being. In their place Mooney (2005) argues that societal resources should be allocated on the basis of community preferences for “what sort of society citizens want, including what sort of social institutions they want and what sort of rules or principles they want to govern these social institutions.” These preferences would determine the

objectives of the healthcare system, and would inform efficiency in terms of what it was the health system was trying to achieve (Mooney 1998b). The better the health system met these objectives – that is, the better societal preferences were satisfied – the greater value society would derive from the healthcare system.

Although citizens may indeed be ill-informed about which specific healthcare interventions should be provided, Mooney (1998b) argues that they can and should contribute to the principles by which healthcare resources are allocated. As noted above, these principles are inherently subjective and do not necessarily require technical expertise. Once these principles are defined, the re-allocation of resources necessary to achieve these equity and efficiency objectives would be left to professional decision makers and clinicians at the meso and micro levels (Mooney 1998b; Nord et al. 1999). The different roles of the public and the decision makers reflects the role of objective knowledge and expertise at different stages of the priority setting process (Buchanan et al. 1998). At the macro level, there is no objectively best allocation of resources; it is a subjective judgement that ultimately rests upon tastes, perspective and persuasion. Once the objectives of the healthcare system have been defined, however, the allocation of resources at the meso/micro level to best meet these objectives is a technical matter that relies on professional knowledge and expertise.

2.5 Societal preferences in priority setting: the equity-weighted QALY

Recent reviews have suggested that society is concerned about factors other than QALY gains, and may be willing to sacrifice aggregate QALY gains to prioritise patients on the basis of characteristics such as age, social role or disease severity (Sassi et al. 2001; Schwappach 2002a; Dolan et al. 2005; Stafinski et al. 2011). Consistent with these suggestions of societal support for equity as well as efficiency in the allocation of health and healthcare resources, operational applications of QALY maximisation have tended to ease the strict QALY maximising decision rule and allow for consideration of equity alongside efficiency in priority setting decisions. For example, health economic evaluation guidelines from the Canadian Agency For Drugs and Technologies in Health (CADTH) note that age, sex, ethnicity, geographic location (usually understood

as remoteness), socioeconomic group or health status may be relevant to some evaluations (Canadian Agency For Drugs and Technologies in Health 2006). Similarly, the National Institute for Health and Care Excellence (NICE) citizens council, which is intended to represent UK public opinion on overarching moral and ethical issues (National Institute for Health and Care Excellence 2013), has noted that factors such as the age of the patient, disease severity, or life-saving treatment may justify greater priority (National Institute for Health and Care Excellence 2008). However, the inclusion of factors other than length and quality of life has tended to be *ad hoc*. There is little specific guidance on when it is or is not necessary to consider these factors, or how they should be weighted relative to each other or to the objective of maximising QALYs. This lack of consistency may jeopardise the public trust in the priority setting process.

One way to explicitly incorporate the distributional preferences of society into priority setting decisions is through what is referred to as the ‘equity-weighted’ QALY. This measure would weight QALY gains to reflect the strength of the equity-efficiency trade-off, and its key feature is that the sum of equity-weighted QALYs accruing to any particular individual can be greater or less than the sum of their unweighted QALYs. Note that at the aggregate level, however, the sum of equity weighted QALYs must equal the sum of unweighted QALYs, and for each patient that receives a greater weight another must necessarily receive lower weight (Ham & Coulter 2001; Wailoo et al. 2009). Culyer (1989) suggests that such a measure, by explicitly integrating equity and efficiency, addresses allocative as well as technical efficiency.

The use of the equity-weighted QALY as a measure of value in healthcare leads to what Nord (1995b) describes as ‘cost-value analysis,’ where the objective is to maximise the total *value* of QALYs gained, rather than the sum of individual QALY gains (Mooney 1998b; Nord et al. 1999). There is nothing in the equity-weighted QALY that requires a democratic or Communitarian approach to defining the relevant equity considerations, but for the reasons discussed above these approaches appear to have advantages for defining equity weights. Such a measure is also consistent with Williams’ (1996) view that “QALYs will also have a role in more complex rules, and more complex rules

will almost certainly be needed if collective priority-setting is to reflect the views of the general public.”

An advantage of the equity-weighted QALY is that it offers an escape from the theoretical ‘QALY trap’ of conventional QALY maximisation, where “the health-related quality-of-life of any health condition determines not only the benefit of curing the condition but also the benefit of saving the life of someone with that condition.” (Ubel et al. 2000) This implies that saving the life of a person with a permanent disability (e.g. paraplegia) is less valuable than saving the life of someone who is otherwise in perfect health, since the person with paraplegia will generate fewer lifetime QALYs.⁶ Conversely, if saving the life of a person with paraplegia is to be considered equally valuable, it is necessary to regard a cure for paraplegia as having no value (Menzel 1999). With an equity-weighted QALY, societal value is not constrained by the individual utility gained, and it is possible to value the two lives equally. Therefore, “the strength of a claim is not a function of an individual's ability to manage to feel harmed. Harms, and the strength of harms, are for the society to judge.” (Mooney 1998b)

The equity-weighted QALY is based on the assumption that efficiency and equity are commensurate concepts, and that efficiency in maximising health gains can be traded off against concerns for equity. As Sassi et al. (2001) note, this implies that a more equitable intervention can be less efficient and still be ranked favourably relative to a more efficient but less equitable intervention. There is a limit to this equity-efficiency trade-off, though, and at some point an equitable but inefficient intervention will be ranked less favourably than a more efficient intervention with a less equitable distribution of benefits. The objective over the remainder of the thesis is to estimate the strength of the equity-efficiency trade-off for different aspects of equity, and the next chapter will review factors that may be relevant to this trade-off.

⁶ Few economic evaluations incorporate utilities at the individual level, so the QALY trap is more of a theoretical than a practical matter, but the implication holds nonetheless.

Chapter 3: Empirical ethics review

A fundamental challenge in the extra-welfarist approach is defining the set of non-utility characteristics relevant to priority setting. Strict theoretical QALY maximisation, based on a view of well-being as health, defines this set solely in terms of length and quality of life, and the number of patients benefitting. More pragmatic applications of this framework have allowed for the expansion of set of relevant factors to include implicit consideration of characteristics such as age or disease severity. However, as noted in the previous chapter, the inclusion of factors other than length and quality of life has tended to be *ad hoc*, with little guidance for when these factors should be considered, or what their weight should be relative to efficiency. In addition, although many of the factors mentioned in the CADTH and NICE guidelines are consistent with recent evidence on societal preferences, the factors included in a particular evaluation, and their relative weights, ultimately reflect decision maker rather than societal preferences. As an alternative, a democratic or Communitarian approach would allow for the set of relevant characteristics, and their relative weights, to be defined by the community (Mooney 1998b; Menzel 1999; Callahan 2003a).

The most straightforward approach to identifying these factors would be simply to ask representative members of a community which attributes or characteristics they consider important. Indeed, such preference surveys have been relatively common in health economics (Nord et al. 1995; Bowling 1996; Mossialos & King 1999). As Mooney (1998b) acknowledges, however, the preferences elicited by such surveys will only be 'good' to the extent that the society from which they derive is also 'good.' That is, society may hold

preferences that are irrational or perverse by, in Mooney's terms, "some universalist principle." Likewise, Ubel et al. (1999) noted that although it is important to consider public preferences in healthcare priority setting, these preferences may not always be fair. Although irrational or perverse preferences may reflect a societal consensus, it is difficult to accept that incorporating community preferences for denying men or particular ethnic minorities healthcare, for example, would improve the moral legitimacy of the resulting priorities. Daniels (1998) goes further, and argues that majority support in a preference survey is not sufficient grounds for distributing something as fundamentally important as healthcare. Such surveys, he suggests, reveal tastes rather than reasons and therefore lack legitimacy: "settling moral disputes simply by aggregating preferences seems to ignore fundamental differences between the nature of values and commitments to them and tastes or preferences." He argues that a deliberative process is required to assure the minority that allocation preferences are based on reasons that they can accept as relevant.

Richardson and McKie (2005) agree that ethically important decisions cannot be resolved by empirical methods alone, but they suggest that deliberation by itself is also insufficient :

The superiority of one theory over another – ethical or otherwise – cannot be determined by logic alone, and yet there must be some agreement about what constitutes a better theory. Neither the discipline of economics nor ethics provides a satisfactory answer to this question.

They propose that "defensible principles for allocating healthcare should be derived in an iterative way, involving both an empirical study of population values and ethical analysis of the results." Richardson (2002) described this process of identifying factors that are *relevant* but also, in some sense, *fair*, as 'empirical ethics.' This process is consistent with Broome's (1989) view of distinguishing claims, which carry some ethical obligation, from the other reasons that an individual may be entitled to some share of a limited societal resource.

Richardson (2002) used the term 'empirical ethics' to describe a process whereby guiding ethical principles are inferred from empirical investigations of societal preferences. That is, evidence of public support should be taken to imply

some normative quality. However, Hausman (2002) has suggested that this represents a form of ‘moral relativism,’ whereby what is morally right or wrong is reduced to social consensus or even a simple majority. He feels such a position is untenable. Slavery, he notes, was once held to be acceptable by a majority of citizens in many countries, but this support did not make it right or ethical. Like Mooney (1998b) and Ubel et al. (1999), however, Richardson recognised that societal preferences should themselves be subject to ethical scrutiny, and that at times it may be necessary for decision-makers to over-ride or ‘launder’ (Goodin 1986) some preferences. The challenge is to identify *which* preferences should be excluded. In this regard, Ubel et al. (1999) proposed that principles for priority setting should reflect “quantitatively significant” societal preferences, and be “consistent with some coherent and defensible ethical theory of justice.” In their view, preferences with a ‘trivial’ impact or minimal support should be excluded, as should preferences that cannot be justified by some coherent theory of justice. This approach, of subjecting potentially relevant attributes to an empirical and an ethical filter, was adopted here to identify a set of attributes that may be considered fair as well as relevant.

To develop this ethical filter, section 3.1 discusses prominent theories of distributive justice that might guide an empirical ethics approach. As the overall objective was to provide specific guidance to the allocation of societal resources, the emphasis was on theories that suggest a specific maximand over those that advocate a particular process. Section 3.2 applies empirical and ethical filters in reviewing the empirical evidence around public support for different attributes or characteristics, and the ethical justifications (or lack thereof) for each attribute based on the theories of justice discussed. Section 3.3 contrasts the attributes identified here with the processes and attributes used by other elicitation studies in this area. Finally, section 3.4 discusses the specific attributes that were judged fair and relevant by this process, as well as some of the limitations of an empirical ethics approach.

3.1 Theories of justice in the allocation of healthcare

As Richardson and McKie (2005) state, “the assertion that one state of the world is better than another is always and unavoidably based upon an ethical theory or belief.” In order to provide an ethical basis for the attributes included among the set of relevant characteristics, this section will briefly review prominent theories of distributive justice in healthcare, and draws heavily on reviews by Williams and Cookson (2000) and Konow (2003). It is important to note that there are often strong criticisms of all the theories identified here, and it is not the intent of the review to argue for an ideal or universal theory of justice. Instead, the different theories will be discussed with respect to their respective visions of how to allocate inevitably scarce healthcare resources.

In this regard, Williams and Cookson (2000) distinguish between theories of distributive justice that specify a specific objective, or ‘maximand’, and those that do not. Theories without a maximand include ‘pure procedural’ theories such as Libertarianism, Contractarianism, Participatory Democracy, and Accountability for Reasonableness, which emphasise the *process* by which a fair outcome is reached, rather than the outcome. Similarly, principles such as the absence of envy, equality of access and rule-of-rescue set ‘side conditions’ to determine whether an outcome is fair, but again do not specify an overall objective. Principlism and the Pareto principle may also arguably be included amongst these side condition principles. Finally, theories with a specific maximand include need, maximisation, egalitarianism, and Rawls’ difference principle.

3.1.1 Pure procedural theories

Among pure procedural theories, *Libertarianism* rejects any role for the government, and in the context of healthcare holds that publicly-provided healthcare should be replaced by private insurance. Any distribution that results from such a free-market arrangement is inherently just (Williams & Cookson 2000; Nozick 1974). Libertarianism does not accept that there is ever a justification for prioritising one patient’s rights over another’s, arguing that such prioritisation would amount to “a utilitarianism of rights.” (Nozick 1974) *Contractarianism* holds that the free and collective agreement of individuals to a

particular arrangement shows that it has some normative property (e.g. legitimate, just, obligating, etc.) (D'Agostino & Gaus 2008). A 'constructivist' interpretation of a social contract views an agreement as normative by virtue of the collective agreement itself, while a weaker 'indicative' interpretation views a collective agreement as evidence of a normative quality, but not a normative justification in itself (D'Agostino & Gaus 2008). Recognise, though, that a social contract is only meaningful under conditions of 'reasonable pluralism'; if all individuals had precisely the same set of preferences, there would be no value in demonstrating that they could agree on something. This pluralism implies that it is extremely unlikely that *all* individuals will ever agree on something, and therefore a social contract is not defined by what people *do* agree to, but rather by what they *would* agree to, if they were all hypothetical 'reasonable individuals,' without biases or false beliefs (D'Agostino & Gaus 2008). In this sense Contractarianism and the social contract are based on the hypothetical agreement of hypothetical individuals – what Dworkin (1989) objected to as a doubly hypothetical agreement. Similar to Contractarianism, *Participatory Democracy* as a theory of distributive justice holds that any distribution arrived at through a fair democratic process is just. Both Contractarianism and Participatory Democracy are similar to Communitarianism, in that all three reflect community preferences. But whereas Contractarianism and Participatory Democracy are based on the idea that community agreement implies a normative property, Communitarianism, as understood in the context of this thesis, simply holds that allocating resources according to community preferences will maximise community welfare; it makes no normative claim about the inherent fairness of such a distribution. Finally, Daniels and Sabin's (2002) '*Accountability for Reasonableness*' (A4R) defines four process conditions to a fair outcome: the publicity of decisions and rationales; the rationale for decisions should be relevant and be acceptable to 'fair-minded people'; there must be an appeals mechanism for challenging and potentially reversing decisions; and the process must be publicly regulated to ensure the first three conditions are met. More generally, Dolan et al. (2007) identify six broad characteristics of procedural justice: a means by which affected or potentially affected parties can have the opportunity to contribute to the decision making

process; neutrality in decision making, or the ability of decision makers to separate themselves from preconceptions and self-interest; consistency in the roles accorded to similar people in the decision process; a mechanism for assessing the accuracy of information to be used in the decision making process; an appeals and reversal process; and transparency in the decision making process.

3.1.2 'Side condition' principles

Among principles that define the characteristics of a fair outcome without specifying a maximand, the *absence of envy principle* defines a fair situation as one where no one envies anyone else, taking into account all aspects of a person's circumstances (Williams & Cookson 2000). Creating such a fair situation requires a compensation principle to adjust for inherent differences between individuals, although this compensation is generally not defined within the theory itself. Within the healthcare context, the absence of envy principle has often led to the idea of equality in initial resources, or resource egalitarianism, which will be discussed in more detail below. *Equality of access* defines fairness as equal access to healthcare, consistent with the concepts of horizontal and vertical equity. Horizontal equity requires similar individuals be treated similarly, while vertical equity requires dissimilar individuals be treated dissimilarly (Culyer 2001b). Williams and Cookson (2000), though, suggest at least four possible interpretations of 'access' in healthcare – the quantity of healthcare utilization (e.g. physician visits); the cost of healthcare utilization; the maximum attainable healthcare; and the opportunity cost of healthcare – and criticise this principle on the grounds that it focuses too narrowly on healthcare as an end in itself, and does not consider an overall objective in terms of health or well-being. The *rule-of-rescue* holds that society has an ethical duty to do everything possible to rescue identifiable individuals from imminent death and is the basis of much of clinical ethics, but does not specify any distribution of resources outside of the single individual (McKie & Richardson 2003). A number of authors, though, argue that it is irrational as well as unfair to devote resources to people who happen to be in immediate distress at the expense of others who may have a greater objective claim to healthcare resources. Thus, it

is difficult to consider the rule-of-rescue a true theory of distributive justice (Williams & Cookson 2000; McKie & Richardson 2003; Hauck et al. 2004). Rather than defining one condition to a fair outcome, *Principlism* defines a set of principles that form the basis of much of modern medical ethics: respect for autonomy, or the right to make one's own decisions; non-maleficence, or the requirement to do no harm; beneficence, or the prevention of harm and the provision of benefit; and justice in the fair distribution of resources (Callahan 2003b; Beauchamp & DeGrazia 2004). While there is no mechanism for resolving conflicts between the principles, Callahan (2003a) suggests that all the other principles can be interpreted as protecting or promoting the autonomy of the individual – any conflict between the principles should be resolved in favour of the outcome that is most consistent with autonomy. Beyond the primacy of individual autonomy in decision making, though, Principlism offers no guidance in how decisions should be made for the benefit of society (Callahan 2003b). Principlism is consistent with many of the principles of neo-classical welfarist economics with its emphasis on individual sovereignty and welfarism, and in this sense, it is largely inconsistent with a Communitarian perspective. Finally, the *Pareto principle* holds that an outcome is fair ('Pareto optimal') if resources cannot be reallocated in such a way that the welfare of at least one member of society is improved without making anyone else worse off (Sugden & Williams 1978). Konow (2003) argues that the Pareto principle has been widely embraced by economics on the grounds that it requires "an ostensibly innocuous value judgement," even though its strict reallocation condition means that the Pareto principle will accept flagrantly unequal distributions as fair if resources cannot be reallocated without creating a 'loser.'

3.1.3 Theories with a specific maximand

The principles discussed above are largely deontological, in that the fairness of an action is judged by its adherence to a particular set of rules or principles. An alternative class of theories of justice are those with a specific maximand. Such theories are consequential in that the fairness of an action is judged solely by the outcomes it generates (Konow 2003; Alexander & Moore

2008). This class of theories includes need principles, maximising principles, and egalitarian principles, including Rawls' Difference principle.

Need principles advocate the distribution of healthcare in proportion to need, consistent with concepts of horizontal and vertical equity, which require that equally 'needy' individuals to receive equal preference over equally 'less needy' individuals, regardless of any other characteristics of those individuals (Hauck et al. 2004). Konow (2003) notes that the need principle requires that a just allocation of resources provide for basic needs equally across individuals, and suggests that this principle tends to dominate when basic needs are endangered. This is consistent with Walzer's (1983) argument that healthcare is a special good which requires a special kind of distributive principle; specifically, that whereas consumer goods can be fairly distributed according to market principles, healthcare should be distributed according to need.

The key requirement to operationalising this principle, though, is an appropriate definition of need. Cookson and Dolan (2000) reject a definition based on 'clinical need', as they suggest such a criterion leads to a procedural principle whereby "any allocation is correct so long as a clinician has taken it." Instead, they identify five potential conceptions of need, each with slightly different implications for healthcare allocation:

- *Need as the degree of immediate threat to life* implies that saving (or prolonging) a life should always take priority over enhancing life.
- *Need as the degree of immediate ill-health* includes immediate threat to life, but also encompasses immediate pain and suffering and implies those in more severe states should take priority.
- *Need as the degree of lifetime ill-health* takes a broader perspective and considers an individual's lifetime health experience. Individuals who have had a relatively long, healthy life would have less priority.
- *Need as the degree of immediate capacity-to-benefit* interprets need as the ability to gain from effective treatment. By this definition, if an individual cannot gain from treatment, they have no need for healthcare. Similarly, Culyer and Wagstaff (1993) argue that the provision of *ineffective* healthcare should not attract any equity concerns, except insofar as it

would be inequitable to use resources that could be used to promote equitable outcomes elsewhere. This principle emphasises health gains without considering duration.

- *Need as the degree of lifetime capacity-to-benefit* expands capacity-to-benefit to include consideration of duration in terms of life expectancy remaining.

Advocates of *maximising principles* take the view that justice is ultimately a matter of maximising the sum total of human happiness. In a health context this implies allocating healthcare so as to bring about the best possible consequences, in terms of aggregate population health, most commonly defined in terms of quality-adjusted life years (QALYs), or something broader, such as well-being or ‘flourishing’ (Cookson & Dolan 2000). Although maximising principles in healthcare are broadly utilitarian, they do not conform to a welfarist definition since well-being is not defined by subjective individual utility but by a more limited conception of health-related utility. Indeed, within healthcare, even broader measures of well-being or flourishing are most often understood in objective terms such as capabilities rather than subjective utility (Culyer 1989; Cookson & Dolan 2000). Maximising principles are the basis of the QALY maximisation approach and correspond closely with an interpretation of need as the lifetime capacity-to-benefit (Culyer 1989; Coast 2009). The key distinction between the maximisation and need principles is that whereas maximising principles would concentrate gains amongst those most able to benefit, possibly to the exclusion of those who could gain less, need principles allocate resources *proportionate* to need, implying at least some resources to those with lesser need (Cookson & Dolan 2000).

Egalitarian principles advocate allocating healthcare so as to reduce inequalities in health. As described by Daniels (1990), egalitarianism is willing “to forego delivering a greater benefit to someone who is already better off in order to deliver a lesser benefit to someone who is worse off.” However, Konow (2003) notes that this relatively simple rule is complicated by different conceptions of equality, and Daniels (1990) identifies at least three potential targets for egalitarian concerns in a healthcare context: equality of *welfare*, equality of *resources* to pursue welfare, and equality of objective *capabilities*.

Equality of welfare can be thought of as an operationalisation of the absence of envy principle, and requires that all individuals be equally happy with their situation in life. However, as each individual's welfare is a function their preferences, an unequal distribution of resources may be required in order to achieve an equal distribution of welfare. As Sen (1985) suggests, individuals differ with respect to their ability to convert resources into well-being, and therefore individuals with 'expensive tastes,' for example, may require a greater share of resources to achieve a given level of welfare. In such cases, Daniels (1990) argues that egalitarian concerns have been hijacked. An alternative interpretation of equality of welfare is offered by Williams (1997), who suggests that every individual is entitled to a certain quantity of lifetime health (i.e. their 'fair innings') and that individuals who have gained a greater share of their entitlement should have a weaker claim to societal healthcare resources. An absolutist interpretation of this argument would hold that there is no value to be gained by treating patients who have achieved their full share of life years or healthy life years and would deny treatment to elderly patients, while a relativist interpretation would give relatively greater priority to younger patients (Tsuchiya 2000). In general, the further an individual is from achieving their fair allotment of healthy life years, the stronger their claim relative to those who have already achieved their fair innings.

Equality of resources, or resource egalitarianism, holds that justice requires each individual to have the same initial resources in order to pursue their welfare but does not prescribe a particular outcome; outcomes are determined by each individual's free choices. In this perspective, poor health is just if an individual had an opportunity for full health but failed to achieve it through their own choices (Cookson & Dolan 2000). However, resource egalitarians also generally hold that circumstances over which individuals have no control should not adversely affect their life prospects. An unequal distribution of resources may therefore be justified in order to compensate individuals disadvantaged by 'brute luck' beyond their control (Daniels 1990; Anderson 1999). Because of this compensation condition, resource egalitarianism is also called 'luck egalitarianism.' (Anderson 1999; Arneson 2000; Feiring 2008) Anderson (1999), though, does not accept that egalitarian

principles can be used to justify fundamental inequalities, no matter what their cause. She rejects luck egalitarianism on the grounds that it fails the most basic test of any egalitarian theory: “that its principles express equal respect and concern for all citizens.” Luck egalitarianism, she argues, effectively dictates what people can do with their freedoms, and abandons individuals judged to have made poor use of that freedom.

Finally, equality of capabilities holds that the objective of healthcare should be to maintain an individual’s “normal opportunity range.” (Daniels 1990) This view sees health as instrumental to an individual’s overall well-being, and that fair equality of opportunity requires that an individual have opportunities equivalent to others *with the same talents and skills*. Like resource egalitarianism, free choices that affect an individual’s range opportunities are not unjust, so equality of capabilities would prioritise healthcare for individuals disadvantaged by brute luck while at the same time limiting healthcare to those with a normal range of opportunity. Daniels (1990) recognises that a particular disease may have a different impact on the normal range of opportunity at different stages of life, and suggests that resources should be allocated so as to protect a *contextual* range of opportunity, thereby contributing to a fair distribution of resources between age groups.

The imperative of maintaining an equal range of opportunity is conceptually similar to Capability theory, which holds that the objective of policy should be to promote and maintain the capabilities necessary to achieve a range of ‘functionings.’ In this context, capabilities are what a person can do, even if they choose not to translate these capabilities into a specific functioning (Cookson 2005; Hausman & McPherson 2006). To illustrate, literacy would be a capability, while reading for pleasure would a functioning (Sen 2011). But although the Capability approach has generally been interpreted as advocating an equal distribution of capabilities (Coast et al. 2008a), Sen (2011) argues that the Capability approach does not prescribe a specific maximand, but rather offers an ‘informational focus’ that society should consider in assessing justice and injustice. In this sense, the Capability approach can be viewed more as the ‘currency’ of distributive justice, similar to the QALY, rather than as a specific theory of distributive justice.

Rawls (1999; 2001) proposed a theory of justice based on two principles: first, 'primary goods', including rights, liberties and opportunities, should be distributed equally and at the maximum level that is compatible with everyone receiving the same allocation. This principle is very close to resource egalitarianism. Second, where there are inequalities, these should be arranged in order to benefit the least advantaged groups. This second principle has become known as *Rawls' Difference Principle*. Although Rawls' theory is based on an equal respect for all persons, his primary concern is for the absolute position of the least advantaged group. If it is possible to improve the absolute position of the least advantaged by having some inequalities, Rawls' Difference principle prescribes inequality up to the point that the absolute position of the least advantaged can no longer be improved (Lamont & Favor 2008). In this sense Rawls argues for a lexicographic welfare function where the absolute position of the least advantaged determines overall societal welfare (Mueller 2003). In justifying his theory, Rawls (1999) imagines an initial position in which people are behind a 'veil of ignorance' and would not know their position in society. In this 'original position', he argues that free and rational individuals would understand that it was equally probable that they could be well-off or badly-off, and would accept a social contract based on equality in order to minimise their risk by ensuring that the worst off are as well-off as possible.

Although Williams and Cookson (2000) have interpreted Rawls' Difference principle in the context of healthcare as prioritising those in the most severe health states, they note that Rawls explicitly excluded health from his list of 'primary goods.' First, Rawls felt that health was distributed by nature as much as society. Second, he felt health is an end in itself, not just a means to pursue other ends. Third, a strict application of the Difference principle could result in excessive share of healthcare resources going to those in the most severe health states. However they also note that Rawls has suggested that the Difference principle may not apply once all members of society have been brought up to a minimum level of health, similar to Daniels' (1990) interpretation of equality of opportunity.

Similar to Rawls' difference principle, as well as to aspects of the rule-of-rescue, *Prioritarianism* holds that the worst-off should have priority over those

that are better-off (Parfit 1997; Arneson 2000). Prioritarianism is distinguished from different forms of egalitarianism by its concern for absolute, rather than relative well-being. To illustrate, a 'fair-innings' egalitarian would assign priority to a moderately ill child and a very ill senior on the basis of their relative accumulation of lifetime health. The child has gained relatively less of her fair-innings and so deserves relatively greater priority than the senior. A Prioritarian, in contrast, would assign priority on the basis of absolute well-being: the senior is more severely ill, and so deserves greater priority. Hausman and McPherson (2006), though, suggest that an emphasis on the worst-off will tend to have the effect of lessening inequalities, making the distinction between Prioritarianism and egalitarian theories relatively insignificant.

Few advocates of egalitarian principles would pursue equity as the sole objective, and instead combine equality with other principles of justice such as need or maximisation (Cookson & Dolan 2000; Culyer 2001b; Hausman & McPherson 2006). In theory, a strictly egalitarian focus on health differences could achieve equality by reducing rather than improving overall health, so to avoid this result, egalitarianism might be combined with maximising principles; what Parfit (1997) refers to as 'pluralist egalitarianism.'

3.1.4 Defensible theories of justice

In considering theories of distributive justice, particularly as applied within healthcare, Williams and Cookson (2000) adopted an economic decision-making perspective and rejected deontological theories and principles on the grounds that they lack a maximand and therefore offer no specific distributional guidance to decision makers. This is particularly true where there is no optimal solution and some trade-off must be made between 'unjust' alternatives (Williams & Cookson 2000; Hausman & McPherson 2006). This pragmatic justification for a decision-making perspective was adopted here, and deontological theories and principles were not accepted as a primary ethical justification for particular preferences or attributes, although they could be acknowledged as secondary considerations.

Instead, need principles, maximising principles, and egalitarian principles, including Rawls' Difference principle and Prioritarianism, were the primary

theories of distributive justice used to support the inclusion of different attributes. The decision to exclude deontological theories and principles should not be interpreted as a reflection of their coherence or defensibility, but simply of their practicality for the specific purposes of this empirical ethics review. It is also important to recognise that this decision introduces an element of subjectivity into the review. Including deontological theories such as Contractarianism or Participatory Democracy would likely have identified a different set of fair and relevant attributes. Indeed, as mentioned in section 3.1.1, those two theories in particular suggest that any distribution based on collective or majority agreement is, by definition, fair. This effectively reduces ‘ethically defensible’ to ‘majority support.’ Although this may be consistent with the idea of inferring ethical principles from population preferences, this relativism is rejected here in favour of Williams and Cookson’s (2000) decision-making perspective.

3.2 Attribute Review

Attributes potentially relevant to a Communitarian approach to priority setting were identified through a review of the health economics, medical, and ethics literature. The review took a ‘citation pearl growing’ strategy, beginning with reviews by Schwappach (2002a) and Dolan et al. (2005). The bibliographies of these reviews were searched and the ‘related articles’ feature of PubMed and Web of Science was used to identify other potentially relevant studies. Keywords from the Schwappach and Dolan reviews were also searched in PubMed, EconLit and Google Scholar. A pearl-growing strategy is suggested to be particularly useful for interdisciplinary topics where relevant studies may use different keywords and be found across different citation databases (Schlosser et al. 2006).

From these results, four reviews, by Sassi et al. (2001), Schwappach (2002b), Dolan et al. (2005), and Stafinski (2011), were deemed comprehensive in that they discussed a broad range of attributes that may be relevant to preferences for health and healthcare. Additionally, a review by Olsen et al. (2003) considered the relevance of three broad categories of personal characteristics: a person’s relations to others, a person’s relations to the cause of

illness and the person’s ‘self’. While they also noted age and aspects directly related to efficiency and distributive justice, such as health gains and severity, they specifically excluded these factors from the discussion as their emphasis was on personal characteristics. Between them, these five reviews identified and discussed 14 unique concepts or factors (see Table 3.1). These factors were taken to represent the set of attributes potentially relevant to a societal perspective on healthcare priority setting.

Table 3.1: Potentially relevant attributes

| Attribute | Sassi et al. (2001) | Schwappach (2002a) | Olsen et al. (2003) | Dolan et al. (2005) | Stafinski et al. (2011) |
|--------------------------|---------------------|--------------------|---------------------|---------------------|-------------------------|
| Age | ✓ | ✓ | ND | ✓ | ✓ |
| Social role/productivity | | ✓ | ✓ | ✓ | ✓ |
| Lifestyle/responsibility | ✓ | ✓ | ✓ | ✓ | ✓ |
| Prior healthcare | | ✓ | | | |
| Social inequality | ✓ | | ✓ | ✓ | ✓ |
| Desert/merit | | | ✓ | | ✓ |
| ‘Self’ | | | ✓ | | ✓ |
| Initial severity | | ✓ | I/ND | ✓ | ✓ |
| Endpoint | | ✓ | | ✓ | ✓ |
| Treatment effect | ✓ | ✓ | I/ND | ✓ | ✓ |
| Duration of benefit | | ✓ | | | |
| Direction of benefit | | ✓ | | | |
| Distribution of gains | | ✓ | | ✓ | |
| Rarity | | | | | I/ND |

I/ND = Identified, but not discussed

Each potentially relevant attribute is discussed in detail below, with an emphasis on the concept the attribute embodies, empirical evidence of public support and the ethical justifications for the concept. Attributes identified in the review that did not have a defensible ethical justification were excluded (or in the term of Goodin (1986), ‘laundered’) from the final set of relevant attributes, and likewise, factors that had a strong ethical justification but limited public support were also excluded as Communitarianism is firmly based on the idea that societal value should *reflect* rather than *impose* preferences.

3.2.1 Age

Preferences for age, or ageism⁷, can be based on a number of ethical principles. Utilitarian ageism is based on a principle of maximising health gains: as younger patients are expected to live longer than older patients, *ceteris paribus*, there is a greater expected value to saving a younger patient (Tsuchiya 1999; Nord et al. 1996). Productivity ageism holds that the very young and the very old have less societal value than individuals at ages in between by virtue of their relative contributions to society. As the very young and the very old tend to require support from the rest of society, while those ages in between tend to be net contributors to society, it may be appropriate to value their health unequally. Indeed, this is the basis of disability-adjusted life years (DALYs) developed by the World Health Organization, which are based on maximising the value of societal productivity (Tsuchiya 2000; Murray & Acharya 1997). A third conception of ageism stems from a perceived moral obligation to save a young life over an older life because they have had fewer life years. This desire to equalise the age at death is known as egalitarian ageism (Tsuchiya 1999; Nord et al. 1996). Williams (1997) takes the egalitarian ageism argument one step further and suggests that it is not age at death that should be equalised, but lifetime health outcomes in the form of QALYs – the so-called ‘fair innings’ argument. Harris (1987; 2005), though, rejects maximising and egalitarian arguments for age-related preferences and argues that healthcare should be allocated so as to maximise lives, based on a position that each life is equally valuable, regardless of its expected length or quality. Such a position denies that there is any justification for considering the age of a patient in determining social value, either explicitly in terms of age in itself, or implicitly, in terms of the expected duration of benefit.

Tsuchiya (1999) reviewed nine empirical studies of age-related preferences and in general the results indicated a consistent preference for younger patients, independent of the age of the respondent. Age-related preference weights consistently declined after middle age and although there was

⁷ ‘Ageism’ in this context is used in the same neutral manner as Tsuchiya et al. (2003), where it simply describes a differential societal value by age rather than implying an unfair discrimination on the basis of age.

some disagreement over whether weights peaked at middle age or childhood, there was no support for equal weightings across all age groups. These preferences were based on a mix of productivity, utilitarian and egalitarian rationale, but when Nord et al. (1996) elicited preferences for pure utilitarian ageism by specifying all patients were the same age and concentrating on duration of benefit, they found evidence of positive but diminishing utilitarian ageism. Respondents favoured younger patients with a greater capacity to benefit, but the strength of these preferences was not proportional to duration of benefit, offering support for a weak version of pure utilitarian ageism.

Studies of hybrid utilitarian ageism combine aspects of utilitarianism and egalitarianism by studying life-saving treatments in patients of different ages. Here as well the preference was to favour younger patients, although again preferences were not proportional to the duration of benefit (Tsuchiya 1999). Support for productivity ageism was mixed. The 'humped-shaped' age-weight profile demonstrated in most of the studies was consistent with productivity ageism, or the view that a year of healthy life is valued differently at different ages (Sassi et al. 2001; Schwappach 2002a; Tsuchiya 1999). However, Busschbach (1993) found that the age-weight profile peaked at the earliest ages (ages 5 and 10), supporting utilitarian and/or egalitarian ageism over productivity ageism. The NICE Social QALY team found a similar result, where a year of full health experienced by a child (aged 0-18) was valued more highly than a year of full health experienced by an adult (Dolan et al. 2008).

Nord et al. (1996) tested the support for 'weak' egalitarian ageism by comparing preferences for younger and older patients with the same capacity to benefit. Weak egalitarian ageism favours the younger patient when both a younger and an older patient can benefit equally, while strong egalitarian ageism favours the younger patient even when the older patient can benefit more. They found that younger patients were consistently preferred when capacity to benefit was equal. Similarly, Baker et al. (2010) found that 64 percent of respondents gave priority to 40-60 year olds over 60-80 year olds, although only 36 percent gave priority to 0-20 years olds over 20-40 year olds, consistent with the humped age profile observed by Tsuchiya (1999) as well as with a productivity ageism

view of priority. They also noted that elderly respondents appeared more likely to prioritise the older age group in both sets of comparisons.

Among the comprehensive reviews, Sassi et al. (2001) concluded that the empirical results demonstrated a preference for prioritising younger patients over older patients, consistent with a view of equity as a concern for equality in lifetime health, although these preferences also appeared to be humped-shaped, suggesting that they may reverse at very young ages. They also suggested that at least some people hold preferences for setting priorities on the basis of the individual value of health and productivity at different ages. Schwappach (2002a) found little support for absolute age cut-offs, but strong preferences for prioritising younger patients over older patients. These preferences exceeded the magnitude that would be expected based on duration of benefit alone, again suggesting a mix of utilitarian and egalitarian preferences. He also highlighted that while there was strong support for prioritising the young, there was much less support for discrimination against the elderly. This suggested significant framing effects in the elicitations – it mattered how the questions were asked. This was also supported by Nord et al. (1996), who found that respondents were reluctant to discriminate between individuals on the basis of age but were comfortable with prioritising budgets for programs that favoured younger patients. Finally, Dolan et al. (2005) found that in most studies respondents gave less weight to older patients, although again it was not clear whether this was for utilitarian or egalitarian reasons.

Green and Gerard (2009) argued that age preferences are confounded by the inability of empirical studies to explicitly separate the effects of age from duration of benefit. In this case, age is primarily a proxy for capacity to benefit and therefore does not represent a true preference for or against specific age groups. Tsuchiya et al. (2003), however, found that although there was some evidence that respondents confused utilitarian and egalitarian motives, there was clearly a humped-shaped age-weight profile, peaking around age 35, once the elicitations explicitly control for the duration of benefit. This finding was supported by Petrou et al. (2013), who elicited the relative value of a fixed health gain across 19 different age groups, from newborn to age 90, and found that value peaked around age 30.

Persad et al. (2009) and Olsen et al. (2003) argued that age is a marker of different stages in every person's lifetime, not a distinct, permanent characteristic that distinguishes one individual from another, and therefore that differentiation by age is not in itself discriminatory. This, along with justifiable maximisation and egalitarian arguments, makes it difficult to conclude that preferences based on age are unfairly discriminatory. Despite Harris' (1987) argument that all lives are equally valuable, regardless of their length, the empirical evidence appeared to demonstrate public support for prioritising younger patients, consistent with utilitarian as well as egalitarian principles.

3.2.2 Social Role & Productivity

Social role refers to the societal duties or responsibilities of an individual. For example, patients with dependents such a young child or an elderly parent, might be considered to play a more valuable societal role than patients without dependents. Similarly, patients with particularly productive skills might be valued more highly than patients with less productive skills. As Schwappach (2002a) pointed out, the hump-shaped age-weight profile discussed above corresponds with values for social roles and productivity. This is not surprising, given the close correlation between social roles, productivity and life stage. However, whereas productivity ageism would discriminate between patients of equal productivity on the basis of age, explicit preferences for social role or productivity would discriminate between patients of equal age on the basis of productivity or prioritise a productive older patient over a less productive younger patient. In this way, preferences based purely on social role or productivity can lead to different allocations than preferences based on productivity ageism, although clearly there is a significant overlap between these concepts.

In their review of the moral relevance of personal characteristics, Olsen et al. (2003) found only limited support for factors related to social role or productivity. Preferences were strongest for patients caring for children or the elderly, although support peaked at 47 percent of respondents to one survey (Olsen et al. 1998). Other surveys reported support for carers ranging between 15 and 33 percent (Olsen et al. 2003), while support for priority based on

productivity was even lower, peaking at 27 percent support for prioritising employed people (coincidentally in the same survey that reported the strongest support for carers). Other surveys of preferences for productivity found little or no support for prioritising breadwinners, employed over unemployed, skilled over unskilled, or teachers over lorry drivers (Olsen et al. 2003).

Preferences based on social roles and productivity can be justified by the maximising principle of the greatest happiness for the greatest number, particularly if the change in utility of all affected parties is considered (Olsen et al. 2003; Mill 1871). In the context of social roles and productivity, one individual's health may have external benefits that increase the welfare of other members of society. An additional QALY to a uniquely productive individual, such as a skilled surgeon for example, may have an aggregate benefit of more than one QALY for society. Olsen et al. (2003) refer to the welfare generated "through caring and personal interaction" as non-pecuniary utility and the welfare generated through what an individual is able to produce as pecuniary utility, and suggest that the aggregate welfare generated through pecuniary and non-pecuniary sources could be substantial. Saving the life of a parent, they noted, generates utility for the patient but also increases the non-pecuniary utility of the child, who benefits from growing up with that parent. Despite an ethical justification based on maximising non-pecuniary and pecuniary externalities, though, there appears to be only limited support for prioritising on the basis of social role or productivity.

3.2.3 Lifestyle and responsibility

A number of the comprehensive reviews noted that society does not appear to be indifferent to a patient's health-related lifestyle and its relationship to the cause of their disease (Schwappach 2002a; Olsen et al. 2003; Dolan et al. 2005). This suggests that society feels more obligated to prioritise patients with 'exogenous' causes of disease over those they feel may have contributed to their disease through unhealthy choices (Olsen et al. 2003). Such preferences are consistent with a luck egalitarian view of the objective of healthcare as offsetting the impact of bad luck that falls on individuals through no fault of their own

(Feiring 2008). Health inequalities that are a result of an individual's own choices are not unjust and thus do not justify priority.

This view appears to be reflected in surveys which have found strong support for prioritising non-smokers over smokers and light drinkers over heavy drinkers (Schwappach 2002a; Olsen et al. 2003). Nord et al. (1995), for example, found that 60 percent of respondents to an Australian survey favoured prioritising non-smokers over smokers, while in the context of liver transplantation, Ratcliffe (2000) found that 71 percent of respondents to a UK survey "agreed or strongly agreed" that preference should be given to patients with naturally occurring liver disease over those with personal responsibility (i.e. heavy alcohol consumption). Dolan et al. (2008) found that respondents showed a statistically significant preference for treating health conditions caused by health service negligence (e.g. MRSA infections) than to conditions where patient lifestyle was a contributing cause. Anand and Wailoo (2000), though, found mixed preferences. Sixty percent of respondents to their UK survey supported prioritising healthcare for individuals infected with HIV through blood transfusions over those infected through illegal intravenous drug use, but only 40 percent favoured prioritising individuals with 'cautious' lifestyles in more general circumstances. A UK choice experiment by Edlin et al. (2012) suggested an even more complex interaction of preferences: although individual responsibility for poorer health prospects tended to be associated with lower priority, the very existence of a health inequality tended to lead to higher priority. The net effect was to give greater priority to patients with poorer health prospects, regardless of the cause of the inequality. Opposition to prioritisation based on individual responsibility was suggested by qualitative discussions conducted as part of the Social Value of a QALY (SVQ) Project (Baker et al. 2010). These discussions drew out the difficulties of assigning blame to individual patients and where the line between culpable and not culpable should be drawn. Sports injuries were mentioned as an example of this difficulty. Finally, in a US survey of public preferences for organ transplantation, Ubel et al. (1999) reported that among respondents who preferred an unequal distribution of scarce organs, only 27 percent preferred lower priority for patients responsible for their disease. It is also worth noting in the context of these findings that respondents who did not

accept prioritising on the basis of a healthy lifestyle were often *strongly* opposed (Schwappach 2002a).

A preference for giving lower priority to patients with an unhealthy lifestyle can be justified by a luck egalitarianism, which holds that unequal health outcomes that are the result of an individual's free choices are just provided that all individuals had the same initial opportunity for lifetime health. The preferences expressed in the surveys noted above, though – particularly those observed by Anand and Wailoo (2000) – appear to be based more on a moralistic attitude against those with an endogenous cause of illness (Schwappach 2002a; Olsen et al. 2003). Nord et al. (1995) and Olsen et al. (2003) refer to this attitude as 'healthism,' or a belief that individuals have a moral obligation to society to live a healthy life. Callahan (2003b) suggests that such an attitude is paternalistic and violates the autonomy component of Principlism, which holds that individuals should live their own lives and make free choices without external coercion or manipulation. Anderson (1999) also finds it difficult to accept that egalitarian principles could be used to justify fundamental inequalities, no matter what their cause.

Olsen et al. (2003) and Feiring (2008) argued that the socioeconomic gradient explains much health-related activity, particularly around smoking and drinking behaviours. The health-related lifestyle of some patients may therefore not have been the result of truly free choices, and they may not have had an equal opportunity for lifetime health. As LeGrand (1987) argued, an individual can only be held blameworthy for those factors substantively within their control. Similarly, Olsen et al. (2003) note that ill health is rarely attributable to one cause and specifically to a person's actions: "one cannot take epidemiological determinants and hold individuals responsible." Together, these arguments suggest that it is difficult to hold an individual solely responsible for their health outcomes.

Although there appears to be at least some support for prioritising patients with a healthy lifestyle, the ethical arguments for supporting such preferences are limited. There may be a maximising justification for prioritising patients with a healthy lifestyle if it is associated with an expectation of better health outcomes, but in general, preferences over lifestyle appear to be motivated by a paternalistic

– or even punitive – application of healthism. Justifications based on luck egalitarianism seem to disregard patient autonomy and basic egalitarian principles of equal respect for all persons.

3.2.4 Prior consumption of healthcare

Schwappach (2002a) was the only author from amongst the five comprehensive reviews to identify the prior consumption of healthcare as a potentially relevant factor. He hypothesised that society may believe that every person is entitled to life saving treatment once in their lifetime, regardless of the cost, but that once a patient has received such a treatment they should ‘step aside’ to allow another patient to benefit. He also hypothesised an alternative position: those patients who require a second life-saving treatment may be viewed as having been ‘betrayed’ by life (analogous to Williams’ fair innings argument) and may therefore deserve greater healthcare priority.

The limited empirical evidence offers some support for the former hypothesis. When asked to reconsider their preferences for saving one of two groups of patients with fatal illnesses after receiving new information on each group’s previous healthcare consumption, 6 percent of respondents to an Australian survey changed their responses to favour the group that had not received prior life-saving treatment (Olsen et al. 1998). Similarly, participants in a UK qualitative study of public preferences for liver transplantation suggested that it seemed unfair to re-transplant one individual while another continues to wait for their first transplant (Wilmot & Ratcliffe 2002).

Ubel et al. (1993), in discussing the ethics of re-transplantation, suggested that preferences for limiting healthcare to those with substantial prior consumption reflect a common sense view of justice where all needy individuals deserve an equal opportunity to benefit from scarce healthcare resources. In this view, individuals should not receive a “second piece of the pie” before some have received their first. However, they argued that such a view is based on a narrow or short-term definition of healthcare and ignores other aspects of health and social spending such as education or primary care. It is not clear that an individual with an episode of substantial healthcare consumption (e.g. a previous

organ transplant) has necessarily consumed an unfair share of overall societal resources when a broader definition is applied.

There may be a maximising justification for considering prior healthcare, but this would only apply to the extent that the quantity of prior healthcare affected future outcomes. A strict application of egalitarian principles might also justify consideration of prior healthcare consumption, although this would require an extreme interpretation that viewed equality in terms of limiting a patient's cumulative access to healthcare. Anderson (1999), in arguing that egalitarianism should be based on principles of inclusion rather than exclusion, appears to reject such an interpretation. In general, lower priority based on a patient's prior consumption of healthcare appears to have only limited evidence of public support, and requires an exclusionary interpretation of egalitarian principles.

3.2.5 Time waited

There may be a preference for those patients who have spent a relatively greater length time waiting for healthcare, reflecting a principle of 'first come, first served.' Such a preference represents a simple, and perhaps simplistic, prioritisation criterion that disregards other factors that may be relevant, particularly an assessment of need.

Many of the relevant empirical studies have been conducted in the context of organ transplantation. A qualitative study of 23 participants in the US found support for consideration of the length of time a patient had been on the wait list when prioritising patients waiting for kidney transplant, although participants tended to mention factors such as the benefit that could be gained from transplant and the consequences of not receiving a transplant before mentioning time on the wait list (Dolan & Shaw 2004). Two conjoint analyses, conducted in the UK (Ratcliffe 2000) and Hong Kong (Chan et al. 2006), reported that time on the wait list was a statistically significant factor in determining the allocation of scarce livers. In the context of appropriate wait times for elective procedures, a survey of 1,101 individuals in Wales, including general practitioners, consultants, health authority commissioners and members of the general public, by Edwards et al. (2003), reported that a majority of

respondents considered factors related to pain and disability as the most relevant attributes, while age, responsibility, ability-to-pay and cost were irrelevant to determining priority. Respondents were mixed regarding time already on the wait list – it was neither clearly relevant nor clearly irrelevant. A substantial proportion of GPs and consultants and commissioners (38-44 percent) felt that priority should be determined by need before time waited, while 32 percent of the general public felt that maximum wait times should be guaranteed, implying that time waited should supersede need after some specific duration.

A strict preference for a ‘first come, first served’ model of prioritisation can be justified by a theory of egalitarianism where all individuals are presumed to be equally deserving in terms of their priority for healthcare, regardless of other characteristics (including need). Indeed, Persad et al. (2009) noted that ‘first come, first served’ is often viewed as an inherently egalitarian form of a natural lottery. However, although such a preference may be superficially consistent with a principle of equality of access, disregarding need in order to prioritise based on time waited would seem to violate the underlying requirement of vertical equity that requires dissimilar individuals (in terms of need) be treated dissimilarly (Culyer 2001b). Persad et al. (2009) also argued that all wait times are not necessarily equal, and that they can be manipulated by individuals with the power, influence or information to get themselves added to a queue sooner. Certainly, where all other relevant factors are equal, principles of egalitarianism and equality of access seem to support the idea that individuals with longer wait times should have some priority. This does not necessarily imply, though, that wait time is itself a relevant factor in prioritisation. Indeed, as Wilmot and Ratcliffe (2002) suggested, a preference based on wait time may simply be a mechanistic criterion that helps avoid, rather than inform, prioritisation decisions.

Although the empirical evidence does not appear to rank wait time above attributes such as need, benefit, or even age, it is clear that there is at least some support for the consideration of wait time in prioritisation. Similarly, while a ‘first come, first served’ approach to prioritisation would appear to offer only simplistic guidance while violating fundamental principles of vertical equity, basic conceptions of fairness would also suggest that longer wait times among

patients of equal need should lead to some priority. Together, this appears to support the consideration of wait time as a factor in healthcare prioritisation. However, such consideration is complicated by the fact that wait times only apply in certain contexts – particularly, as the empirical evidence reflects, elective services and the treatment of chronic conditions. Wait time is not a relevant factor in the context acute services, where need and capacity to benefit are the primary considerations. In view of this restricted applicability, its equivocal empirical evidence and its limited ethical justification, it is difficult to view wait time as a fundamentally relevant attribute in the allocation of healthcare resources.

3.2.6 Societal inequality

In circumstances of societal inequality, there may be a desire to use of healthcare as a tool of social policy. Specifically, a preferential allocation of healthcare resources may be used to compensate individuals disadvantaged in other, non-health aspects of society, most commonly in terms of socio-economic status (SES) (Sassi et al. 2001; Olsen et al. 2003). To the extent that low SES is associated with low productivity, this is the opposite of the desire embodied by greater priority for productivity (e.g. productivity ageism, pecuniary utilitarianism) and reflects a desire to compensate rather than penalise low productivity groups. There is also an interpretation that suggests giving lower priority to high SES groups may be justified on the grounds that they are more able to provide for themselves and have less need for societal resources (Baker et al. 2010).

A review by Olsen et al. (2003) found some support for discriminating based on SES, but none of the included studies demonstrated majority support. Mooney et al. (1995) found that 41 percent of respondents to an Australian survey favoured prioritising low SES groups. A survey of Swedish politicians found a willingness to sacrifice efficiency in order to equalise outcomes between ‘blue collar’ and ‘white collar’ workers (Lindholm et al. 1998), although Sassi et al. (2001) argued that this study may represent a preference for equality between groups rather than a preference for lower SES groups *per se*. Finally, Dolan et al. (1999) found that 23 percent of participants favoured lower priority for rich

groups, 10 percent favoured higher priority for poor groups, 8 percent favoured higher priority for low education groups and 3 percent favoured higher priority for the unemployed. Thirty-three percent also favoured lower priority for individuals with private health insurance, suggesting that respondents may have given lower priority to individuals they felt could 'pay their own way' in the health system.

Olsen et al. (2003) suggested that preferences based on societal inequality may be justified on egalitarian grounds. They distinguished between *general* egalitarianism, which favours an equal distribution of 'well-being', and *specific* egalitarianism, which focuses on one aspect of well-being – in this case, health. Preferences for favouring low SES groups may be justified by specific egalitarianism to the extent that low SES groups are also disadvantaged in terms of health. In this circumstance, greater priority for low SES in the allocation of healthcare may reduce such health inequalities. Such preferences may also be justified by general egalitarianism if the overall well-being of low SES groups can be improved through preferential healthcare allocations. As discussed above, prioritising low SES groups is consistent with an egalitarian desire to equalise the opportunity for lifetime health, particularly if low SES groups suffer from a systemic lack of opportunity. General egalitarianism may also justify lower priority for high SES groups if it narrows the gap in overall well-being, although this would require an extreme interpretation of egalitarianism that was indifferent to an increase or decrease in overall well-being and focused only on a goal of equality. Finally, an alternative motivation for giving lower priority to high SES groups may be based on a desire to, in effect, expand the healthcare budget by requiring those groups that are able to pay for their own healthcare to do so, although this would require ignoring their contributions to the public healthcare system through taxes.

The prioritisation of low SES groups may be justified by both general and specific egalitarian arguments, but there is no evidence of strong public support for such a preference. Although there is also some support for giving lower priority to high SES groups, this appears to be a minority opinion with no clear ethical justification.

3.2.7 Desert and merit

Preferences based on desert or merit reflect the idea that an individual's meritorious or honourable past actions make them more deserving of healthcare, while criminal or dishonourable actions make them less deserving (Olsen et al. 2003). Such preferences are based entirely on retrospective, non-health concerns and do not take into account past, current or future health needs.

Olsen et al. (2003) found little public support for preferences based on desert or merit, with two studies reporting support of less than 5 percent for prioritising patients who have 'contributed a lot to the community.' Ubel et al. (1999) reported that 15 percent of respondents preferred to give intravenous drug users lower priority for organ transplantation, and qualitative interviews suggested that these preferences were based on the perceived merit of drug users and made no reference to the cause of their disease or their relative prognosis. There also appeared to be a convergence of preferences around desert and for a healthy lifestyle. As mentioned earlier, Anand and Wailoo (2000) found that only 40 percent of respondents supported prioritising patients with a healthy lifestyle over those with a more risky lifestyle, but when presented with a more specific choice between patients who developed HIV through a blood transfusion or through illegal drug use, the proportion jumped to 60 percent. This suggested that some categories of risky behaviour were felt to be more acceptable than others. In these last two cases, respondents appeared to be punishing illegal behaviour by giving patients lower priority for healthcare.

There is some precedent for prioritisation on the basis of desert – Olsen et al. (2003) noted the example of separate healthcare facilities for war veterans. They suggested that priority on the basis of meritorious actions may be justified where health needs are a direct consequence of trying to improve the overall well-being of society, and where such actions were voluntary, on the presumption that voluntary sacrifices are more meritorious than paid ones. On the whole, though, they found it difficult to justify priority on the basis of desert, as such an arrangement implies that the healthcare system should function as an "omnipotent Supreme Court" in imposing rewards or punishments. With respect to lower priority for those with past criminal actions, they argued that once atonement has been made through the legal system, a criminal becomes a

free citizen, with all the entitlements to public services of other citizens. In general, there appeared to be little empirical or ethical support for priority on the basis of desert or merit.

3.2.8 'Self'

According to the definition provided by Olsen et al. (2003), 'self' refers to characteristics that are embodied within a person: physical, intellectual or attitudinal. In their review of personal characteristics in setting health priorities, 'self' included sex/gender, race and sexual orientation.

The only evidence in support of priority setting on the basis of such characteristics came from Dolan et al. (1999), who found that 3 percent of respondents favoured higher priority for men, 3 percent favoured higher priority for women and 10 percent favoured lower priority for homosexuals. This suggested a lack of support for these arguably prejudicial preferences, although it must be noted that even 10 percent support highlights the potential pitfalls of directly incorporating public preferences into healthcare priority setting.

Olsen et al. (2003) were unable to identify any ethical arguments to justify higher or lower priority on the basis of any of these characteristics. Rather, they concluded that such characteristics are most likely to be associated with different types of prejudice or bias such as sexism, racism or homophobia. This, along with an absence of empirical support, appears to justify laundering such preferences.

3.2.9 *Initial severity*

It is broadly accepted that healthcare should be allocated according to some definition of need. QALY maximisation has conventionally defined need in terms of an individual's capacity-to-benefit from healthcare, but need in terms of severity of illness or disability has increasingly come to be regarded as a legitimate equity concern (Sassi et al. 2001). These two definitions of need, though, are often at odds with one another. The most severely ill patients – particularly when initial severity is defined by proximity to death – will often have the least capacity-to-benefit, while the less severely ill may tend to have a

relatively greater capacity-to-benefit in terms of life expectancy and expected QALY gains (Culyer 2001b; Hauck et al. 2004; Oliver et al. 2004).⁸

The Norwegian National Health Service has concluded that severity should be of primary importance in prioritising patients, and several other countries, including Finland, France, Germany, Spain, Sweden, and the Netherlands, explicitly consider severity in reimbursement decisions (Shah 2009). A review of 19 empirical studies by Shah (2009) found broad evidence that respondents preferred a health gain to patients starting at a lower point on a quality scale over an equal gain to patients starting a higher point. Indeed, in many of these studies, including Damschroder et al. (2005) and Green (2009), respondents preferred a smaller gain to more severe patients over a larger gain to less severe patients. Similarly, Dolan et al. (2008) found that there was a premium on health gains in the lower half of the quality scale. Using an alternative interpretation of severity, Ubel and Lowenstein (1996) found evidence that respondents were unwilling to prioritise against patients with a poorer prognosis, although the strength of this preference declined as the prognostic differences became larger. Dolan and Tsuchiya (2005), in comparing the relative strength of concerns for the young versus the severely ill, reported a contradictory result. They found that age was a dominant preference, in that respondents preferred to prioritise the young over the old, regardless of the relative differences in life expectancy remaining. Shah (2009) argued that this result highlighted the limited perspective of many studies included in his review, as most focused exclusively on the trade-offs between health maximization and concern for severity, and thus may have failed to capture respondent concerns for other factors.

Life-saving interventions may represent a special case within severity. A number of authors suggested that there is a particular preference for life saving interventions, beyond what would be expected on the basis of preferences for severity (Nord 1996; Wiseman et al. 2003), and a UK review concluded there

⁸ Interpreting severity in terms of proximity to death reduces this example to something of a tautology, as severity implies a relative lack of capacity-to-benefit, but the example holds nonetheless. Patients initially near death may reasonably be expected to have a shorter remaining life expectancy, even with treatment, relative to patients in less severe initial health states.

was a strong willingness to pay for costly life-saving interventions over more cost-effective quality-improving interventions (Shickle 1997). This preference was particularly strong in the case of life-saving interventions for children. There may also be a particular concern for patients at the end of life, defined by NICE as patients with a life expectancy of less than 24 months (National Institute for Health and Clinical Excellence 2009). In this context, severity can be understood as proximity to death. A NICE consultation found that 63 percent of participants supported giving greater priority to patients with a terminal illness and a short (<24 months) life expectancy, although this support was much stronger among the public, patients and carers than among healthcare professionals (National Institute for Health and Clinical Excellence 2009). In contrast, a discrete choice experiment by Shah et al. (2012) concluded that life expectancy was not a driving factor in respondent choices.

Schwappach (2002a) suggested that the desire to prioritise the most severely ill can be interpreted as a variant of the rule-of-rescue, or the imperative that people feel to rescue identifiable individuals from death. McKie and Richardson (2003), though, disagreed. They argued that an emphasis on 'identifiable individuals' distinguishes the rule of rescue from a more general preference to help the worst off. Instead, a desire to prioritise those in more severe conditions appears to be more consistent with Prioritarianism, and the principle of need as the degree of immediate threat to life or ill health. It may also reflect a desire to minimise the differences in well-being between the best and worst off, consistent with Rawls' Difference principle (Rawls 1999) as well as Daniels' equality of opportunity principle (Daniels 1990; Daniels 2001). According to Daniels (2001), the purpose of healthcare is to maintain an individual's normal functioning, thereby protecting their "equality of opportunity." In this view, severity represents the relative impairment of an individual's normal functioning, and the more restricted an individual's range of functioning, the greater their need for healthcare. This was echoed by Doyal (1995), who argued that "the greater the disability caused by illness, the greater the moral entitlement to effective treatment." It is important to recognise, however, that in the absence of an effective treatment, need cannot be said to exist (Hurley 1998), and on this basis, Culyer and Wagstaff (1993) suggested that

severity in itself is not sufficient to justify specific equity concerns. Any conception of priority on the basis of severity must therefore also consider the availability of effective treatment, as NICE (2009) did in limiting consideration of end-of-life priority to those situations where there was also a treatment that could extend life for at least 3 months.

3.2.10 Final health state

QALY maximisation is concerned with absolute health gain, implying that an improvement in health-related utility from 0.1 to 0.3 is equally as valuable as an improvement from 0.6 to 0.8. But while there is evidence that society may be willing to prioritise patients in the most severe health states out of a concern for the worst off, there is also evidence that society may be unwilling to allocate resources to treatments that leave patients in relatively poor health states. This highlights the tension between the interpretation of need as initial severity and need as capacity-to-benefit.

Roberts et al. (1999) found that respondents were reluctant to allocate resources for patients that would remain in a severe health state following treatment, even when such an allocation maximised expected QALYs. In addition, contradictory to evidence showing a preference to treat the more severely ill, Dolan and Green (1998) found that respondents preferred to give treatment to patients in a less severe initial health state and surmised that respondents were concerned about the value of the *post-treatment* health state. Qualitative work by Dolan and Cookson (2000) may reconcile this apparent inconsistency in finding that respondents tended to evaluate health gains in terms of the final health state rather than the relative or absolute improvement. It appeared that treatment must result in some minimum, or threshold, level of quality in the post-treatment health state in order to justify treatment, regardless of initial severity or relative health gain. Results from other authors appeared to support this interpretation. The SVQ Research Team (Baker et al. 2010) found that although 58 percent of respondents preferred to give priority to patients who could move from 60 to 80 percent of full health over those that could move from 80 to 100 percent, only 38 percent of respondents preferred the more severe group when the choice was between a move from 0 to 20 percent or 20 to 40

percent. Abellan-Perpiñan and Pinto-Prades (1999), using constant-sum paired comparison methods, found that although a better final health state was not necessarily a decisive factor in the allocation of resources, respondents were not indifferent to final health state. This suggests that although final health state becomes less important once some minimum quality threshold is reached, the public is willing to discriminate if a patient is not likely to achieve this threshold.

Priority to those likely to finish treatment in a better final health state would likely be opposed by egalitarians and prioritarians on the grounds that this may exacerbate health inequalities and effectively abandon the worst-off. However, a preference for some minimum final health state might be justified by a maximisation interpretation of Daniels' view of 'equality of opportunity', which holds that the purpose of healthcare is to maintain an individual's normal functioning (Daniels 2001). If, after effective treatment, an individual would still be unable to achieve minimum normal functioning, it may be preferable from a maximising perspective to concentrate scarce resources on those individuals that could achieve normal function. The capability approach might be interpreted in a similar maximising context, particularly as Sen (2011) denies that the approach prescribes equality of capabilities. There appears to be empirical evidence and at least some ethical justification for the relevance of final health state in priority setting.

3.2.11 Size of health effect

QALY maximisation implies that when faced with a choice between two patients, priority should go to the patient that can generate the greatest aggregate health gains (as measured by the QALY). However, Ubel et al. (2000) suggested that this principle also implies that if two similar patients with the same condition can be cured, but one patient can be returned to full health while the other will be returned to less than full health as a result of some pre-existing chronic condition, priority should go to the patient with the greatest potential health gain, regardless of their similarity in all other respects. The emphasis on maximising QALYs means that it is less valuable (or even a liability) to cure the

patient with a disability, as they will generate relatively fewer lifetime QALYs.⁹ Harris (1987) suggested that this represents a form of ‘double jeopardy,’ as because of the pre-existing chronic condition, the second patient receives lower priority for their current, unrelated illness. He argues that each life should be regarded as equally valuable, regardless of its relative length or quality.

The comprehensive reviews by Schwappach (2002a) and Dolan et al. (2005) found substantial evidence that the public does not favour prioritisation on the basis of potential health gains, but prefers to give equal priority to individual regardless of their capacity to benefit. A survey by Nord et al. (1995) found that respondents had no preference for prioritising those that could be helped the most over equal priority for all patients. Analogous to the example above, Ubel et al. (1999) elicited preferences for life-saving treatments over two groups: one group had pre-existing paraplegia and could not be returned to full health, while the second group was otherwise healthy and could be returned to full health. Respondents viewed life-saving treatment to be equally important in both groups. A similar result was found by Damschroder et al. (2005). As mentioned in the discussion of severity, a substantial proportion of respondents to a number of surveys even preferred to give priority to patients with the *poorer* prospects, over those that could gain more (Damschroder et al. 2005; Green 2009). Linley and Hughes (2012), though, found evidence of a statistically significant preference for patients that would gain a considerable improvement in health over those that would gain relatively little. The SVQ Research Team (Baker et al. 2010) addressed a different aspect of this issue by estimating the relative value of equal improvements in health-related quality from different points on a quality scale. They found that an improvement from 20 percent to 40 percent of full health was associated with greater value than the same sized improvement from 0, 40, 60 or 80 percent of full health. This highlights the

⁹ This example assumes a multiplicative utility function, as is common in many health economic evaluations. If initially patient 1 is at 100% of full health but patient 2 is at 90% as the result of a chronic condition, and both develop an illness that would reduce their health by 50% for 10 years, curing patient 1 generates 5.0 QALYs $[(1.0-(1.0 \times 0.5)) \times 10]$ and curing patient 2 generates 4.5 QALYs $[(0.9-(0.9 \times 0.5)) \times 10]$. If utility is additive, however, and the illness would reduce both patients’ utility by an absolute 0.5 for 10 years, then the benefit of a cure would be 5.0 QALYs for both patients (0.5×10) .

importance of the context of health improvements, in terms of initial and final quality, over the absolute size of the gain.

Prioritising absolute health gains clearly reflects maximising and utilitarian principles. Indeed, to the extent that health is viewed as intrinsically good or fundamental to well-being, utilitarianism views the maximisation of health as a moral *obligation* (Hausman & McPherson 2006). However, as Anand and Wailoo (2000) noted, the evidence suggests a general belief that it is individuals, rather than the health gains they can produce, that should be treated equally. This belief appears consistent with Harris' (1987) argument of the equal worth of all lives, regardless of their absolute health potential. Finally, Menzel et al. (1999) referred to the 'maintenance of hope,' or the idea that all patients deserve at least the hope of a health gain, not just those that can benefit the most. This, together with empirical evidence showing little support for prioritisation on the basis of absolute improvement, suggested that that absolute health effect was not a primary concern in allocating societal healthcare resources.

3.2.12 Duration of benefit

QALY maximisation assumes that the societal value of health gains is a linear function of the absolute health gain and the duration of benefit: as duration of benefit increases, societal value increases at a proportional rate (Bryan et al. 2002; Dolan et al. 2005). Furthermore, it assumes that quality and duration are 'mutually utility independent.' This means that the preference for a particular health state does not depend on the duration of that state, and that there is a constant proportional trade-off in the proportion of life years that an individual is willing to give up in return for an improvement in quality, regardless of the absolute number of life years involved (Bleichrodt & Pinto 2006). Schwappach (2002a), though, argued that the societal value associated with a particular duration of health benefit is a complex mixture of life expectancy and preferences for age, severity and time, and that it is difficult to disentangle preferences for duration alone.

The empirical evidence appeared to support Schwappach's argument. A qualitative study by the SVQ research team (Baker et al. 2010) suggested an interaction between duration and quality, in that respondents would not want to

live longer in a poor health stage. Another qualitative study of public perceptions of distributive justice in the context of liver transplantation found that respondents were relatively uninterested in differences in survival gains between patients, on that grounds that even a minimal survival gain was important (Wilmot & Ratcliffe 2002). Nord et al. (1996) found support for utilitarian ageism based on a preference for a greater duration of health benefit, but that these preferences were not proportional to duration of benefit: doubling the duration of health benefit did not double the societal value. A study by Dolan and Cookson (2000) also suggested that respondents were more willing to trade-off health gains for other objectives once the number of life years gained exceeded a certain threshold. Together these studies suggested declining marginal value in the duration of health benefit.

Nord et al. (1996) dismissed the idea that discounting in economic evaluations adjusts for declining marginal value in duration in arguing that although discounting reduces the present value of future benefits, it does so to reflect a time preference for benefits occurring now compared to benefits occurring in the future. This is not the same as accounting for a diminishing marginal value of duration. Gafni (1995) used the following example:

A 'first year benefit' occurring 10 years ahead is discounted at the same rate as the last year of a health effect starting in the present and lasting for 10 years. In contrast, a decreasing marginal value based on diminishing returns in respect of quantity would result in a higher value attached to the first year benefit occurring in 10 years than to the last of a 10 year benefit scenario.

It is clear from this example that the marginal societal value of additional life year is not the same, nor even the same concept, as the discounted value of a life year occurring in the future.

A preference for a longer duration of health gain over a shorter duration can be justified by maximisation principles: more years of life are preferred to fewer years of life. However, Harris (1985) argued that an individual with a short life expectancy can place the same value on their remaining time as an individual with a much longer life expectancy, "precisely because it is all the time left." A preference for patients with a longer duration of benefit may also

tend to exacerbate outcome inequalities, contrary to egalitarian principles. These contradictory results seem to support Schwappach's argument regarding the complexity identifying a preference for duration alone, but also suggest that it may be relevant to societal preferences, particularly in its interaction with attributes such as quality.

3.2.13 Direction of benefit

Schwappach (2002a) suggested that the direction of benefit may be relevant to society in terms of a preference for acute or preventive care. Acute care would improve health (i.e. an upward movement on the quality scale or an increase in the duration of health), while preventive care would prevent health declines. Expected utility theory suggests that a gain of 0.5 QALYs should be valued equally to preventing a loss of 0.5 QALYs; thus society should be indifferent between acute or preventative care (Feldman & Serrano 2006). With Prospect theory, however, Tversky and Kahneman (1986) propose that individuals are more sensitive to losses than they are to gains, and that the disutility associated with a loss may be greater than the utility associated with an equal gain. If Prospect theory holds in the context of health, society may indeed prefer preventive over acute care interventions.

The evidence for preferences around the direction of health benefit appeared inconclusive. Three studies showed at least some preference for preventive services. A survey of the public by the British Medical Association and the King's Fund ranked childhood immunisation and screening for breast cancer as the top two priorities from a list of 10 services, ahead of heart transplants, hip replacements and cancer treatment for smokers, suggesting a preference for preventive services (Shickle 1997). Johannesson and Johannesson (1997) conducted a person trade-off exercise comparing preferences for lives saved through preventative care and lives saved through acute care, and found that a life saved through preventive care was valued slightly more highly, equal to 1.2 to 1.4 lives saved through acute care. Finally, Ubel et al. (1998) asked respondents to choose between an intervention that would improve function and an intervention that would prevent further decline, where both interventions had the same absolute magnitude of benefit. They found broader support for

prevention, although the preferences for prevention versus cure were not significantly different when strength of preference was taken into account. Other studies, though, have been more equivocal. A prioritisation ranking exercise for the UK Office of Population Censuses and Surveys (OPCS) found that 'preventative screening and immunisations' was ranked third behind life-saving treatment for children and special care and pain relief for the dying, but ahead of items such as hip replacement surgery (rank 4) and organ transplants and other life-saving surgeries (rank 7) (Bowling 1996). A similar ranking exercise conducted by the City and Hackney Health Authority found similar results, but although preventive services were still ranked behind life-saving treatment for children and special care and pain relief for the dying, it was also ranked below organ transplants and other life-saving surgeries (Shickle 1997). Finally, a German survey found that respondents strongly favoured improvements in health over the prevention of declines (Schwappach 2002b). Schwappach (2002b) suggested that part of the reluctance to prioritise preventive care may lie in the uncertainty around its effect: it is impossible to know for certain which patients will decline in the absence of preventive care, while it is relatively straightforward to identify which patients can benefit from acute care.

A preference for preventive care would generally favour interventions directed toward the healthy rather than the ill, and would seem contrary to Daniel's (2001) and Doyal's (1995) arguments that an individual's relative need for healthcare should reflect the severity of their health state. To the extent that preventive care implicitly or explicitly favours those who have more health to lose, a preference for preventative care would seem to discriminate against more severely ill patients, although this is consistent with strictly consequential maximisation principles. It is important to note that a preference for preventive care on the basis of perceived cost efficiency is not the same as a preference for a particular direction of benefit.

The distinction between acute and preventive care, though, may be largely arbitrary. For example, do life-saving treatments improve health or prevent death? The direction of benefit may simply lie in the timing – a hip replacement in a patient with full mobility *prevents* a deterioration in health-related quality; a hip replacement in a patient with limited mobility *improves*

health-related quality. Given the difficulty of defining precisely what distinguishes preventive care from acute care, it is difficult to interpret the studies presented above. If anything, they appear to demonstrate support for interventions in children more so than a preference for any particular direction of benefit.

3.2.14 Distribution of health gains

QALY maximisation, based on a foundation of consequential maximisation and the potential Pareto criterion, is indifferent to the distribution of health gains: provided that the aggregate gains are the same, large gains to the few are equally valuable as small gains to the many. Society, though, may have a preference for one distribution or the other, independent of the characteristics of the patients or the interventions (Schwappach 2002a; Dolan et al. 2005).

Choudhry et al. (1997) found that 56 percent of health ministry officials in Ontario, Canada preferred a large increase in life expectancy for the few over small gains for the many. Olsen (2000) found a contradictory result, as a clear majority of respondents to a Norwegian survey of the general public preferred a more equal distribution of health gains to maximising health gains. Olsen also suggested that there may be a threshold level for health gains, below which respondents prefer to concentrate gains and above which respondents prefer to distribute gains widely. Rodriguez-Migueza and Pinto-Prades (2002) found a similar result in their survey of Spanish undergraduate students, where respondents preferred to distribute smaller gains to a larger number, provided gains were sufficiently large. The threshold for distributing gains appeared to be around nine additional life years; below nine years, respondents concentrated gains and preferred to give eight additional years to one patient rather than one additional year to eight patients. This threshold effect, as well as differences in preferences between health officials and the general public, may explain the contradictory findings between Choudhry and the other two surveys. Finally, Ubel et al. (1996) asked respondents to choose between two hypothetical screening tests. The first test could screen the entire population and save 1,000 lives. The second, more effective test could only screen 50 percent of the population but could save 1,100 lives. Fifty-six percent of respondents from the

general public preferred to make the less effective test available to everyone. In general, there appeared to be consistent public support for egalitarianism in the distribution of health gains, while the study by Ubel et al. (1996) also demonstrated a clear preference for equality of access over efficiency in saving lives. It could also be argued that this study supports outcome egalitarianism in preferring that everyone has the same, albeit less effective, opportunity to have their life saved, rather than concentrating a more effective opportunity within half of the population, as well as an aversion to an extreme distribution of resources, where half the population receives nothing.

As noted, QALY maximisation is indifferent to the distribution of gains, so long as aggregate gains are maximised. Different egalitarian principles, though, would justify different distributions of a fixed gain. Tsuchiya and Dolan (2009) distinguish between gain egalitarianism and outcome egalitarianism. Gain egalitarianism prefers equality in health gains, suggesting a preference for smaller benefits to the many regardless of their current level of health, while outcome egalitarianism prefers equality in overall health, suggesting a preference for larger gains concentrated among those that are most deprived. A third egalitarian interpretation, based on equality of access, rejects prioritisation on the basis of gains or outcomes, and prefers equal priority to all (Persad et al. 2009). With the exception of the study by Choudhry, the empirical evidence appeared to indicate a preference for more equal distributions of healthcare resources and health gains.

3.2.15 Disease rarity

Related to the issue of the distribution of benefits to the many or the few is the issue of rarity, or the prevalence of a specific disease in the population. Diseases with very low prevalence, usually in the range of 2.5 to 7 cases per 10,000, are defined as 'orphan' diseases (Hughes et al. 2005). Because of the small patient populations, the costs of drug development for such disease can be very high, and the cost-effectiveness of such drugs is often much higher than would generally be accepted (Desser et al. 2010). It is argued that this makes it more difficult for patients with rare diseases to access potentially beneficial drugs, leaving them at a disadvantage relative to patients with more common

diseases (Hughes et al. 2005). The issue in terms of societal preferences is whether the relative rarity of a condition should lead to special consideration in terms of priority and acceptable cost-effectiveness.

Empirical evidence of societal preferences around orphan diseases is limited. A NICE Citizen's Council reported that 16 of 27 members felt that the NHS should, under certain conditions, consider paying 'premium prices' for drugs to treat rare diseases (National Institute for Health and Clinical Excellence 2004). A further 4 members felt that the NHS should pay premium prices for drugs to treat rare diseases under *any* conditions. The remaining seven members felt that funding decisions for orphan drug should be conducted within the same cost-effectiveness framework as any other drug. In contrast, a conjoint analysis of 1,547 respondents in Norway found no societal preference for rarity (Desser et al. 2010). Given a choice between treating an equal number of patients with a rare disease or a common disease, assuming both diseases were equally costly, 70 percent of respondents were indifferent, 20 percent favoured the common disease and only 10 percent favoured the rare disease. This was consistent with the hypothesis that there was no explicit preference for rarity, *per se*. When, in a second scenario, the cost of the rare disease was assumed to be four times more expensive than the common disease, the proportion of respondents favouring the rare disease declined and many of the previously indifferent respondents shifted to favouring the common disease: 47 percent were indifferent, 45 percent favoured the common disease and only 7 percent favoured the rare disease. The authors argued that the relatively high proportion of indifferent respondents in this high-cost scenario reflects a confounding effect of a general concern for fairness and equality in the allocation of healthcare resources rather than true indifference between higher cost rare diseases and lower cost common diseases. Similarly, a conjoint analysis of 213 respondents in Ontario, Canada also found no willingness to pay more for drugs to treat rare disease, or to pay more for each life year gained by a patient with a rare disease (Mentzakis et al. 2011). Instead, respondents gave the greatest weight to severity and treatment effectiveness. Finally, a qualitative Israeli study of 130 individuals appeared to take a middle position: only a minority of respondents favoured prioritising very costly medications for small numbers of patients with rare diseases, while the majority

favoured prioritising medium cost drugs that may be beyond the reach of most patients but that could also benefit a relatively large number (Guttman et al. 2008).

Ethical arguments in support of special consideration for orphan diseases tend to revolve around rights-based arguments that hold “that patients suffering from a rare condition should be entitled to the same quality of treatment as other patients” (Hughes et al. 2005), and a principle of non-abandonment in the allocation of scarce healthcare resources, even where orphan drugs do not meet conventional cost-effectiveness thresholds (Gericke et al. 2005). These arguments are largely compatible with Daniels’ (2001) principle of equality of opportunity, where all individuals are entitled to healthcare necessary in order to maintain a minimum level of normal functioning. However, as Hughes et al. (2005) note, an emphasis on equality-based arguments neglects the fact that decisions that favour higher-cost orphan diseases imply that patients with more common diseases, and who could benefit equally, are less worthy of treatment. As McCabe et al. (2006) argue, priority “for no other reason than rarity of the condition seems unsustainable and incompatible with other equity principles and theories of justice.”

3.3 Attributes in other stated preference elicitations

A summary of the attributes included in other preference elicitations over equity and efficiency in health, and the processes used to identify these attributes, is shown in Table 3.2:

Table 3.2: Attributes in recent stated preference elicitations

| Study | Attribute selection process |
|--|--|
| Ubel & Loewenstein (1996) | Attribute was specific to objective – how do people choose to distribute scarce organs by prognosis? Recipients were specified to be children to avoid considerations of social worth, ability to pay and personal responsibility for illness. Attributes: Probability of survival |
| Abellan-Perpinan & Pinto-Prades (1999) | Attributes were specific to objective – how does priority change with potential for health? Attributes: Relative cost, final health state |
| Ratcliffe (2000) | Attributes selected by investigator to “reflect key decision criteria which |

| | |
|----------------------------|---|
| | <p>respondents may choose to apply in discriminating between potential recipients for donor organs.”</p> <p>Attributes: Age, alcoholic liver disease (responsibility), expected survival, time spent waiting, re-transplant status</p> |
| Bryan et al. (2002) | <p>Attributes selected by investigators “to allow investigation of the core components of the QALY-maximisation model.”</p> <p>Attributes: Number of people, chance of success, survival and quality with treatment</p> |
| Schwappach (2003) | <p>Attributes were chosen by investigator to test preferences for the allocation of healthcare resources. Identification process not specified.</p> <p>Attributes: Healthy lifestyle, Socio-economic status, age, life year gain, final health state, prior life-saving treatment</p> |
| Baltussen et al. (2006) | <p>Attributes selected on “basis of a review of priority-setting criteria..., plus discussion with a range of stakeholders and policy makers.”</p> <p>Attributes: Cost-effectiveness, poverty reduction, age, severity, health benefit, budget impact</p> |
| Chan (2006) | <p>Replicated Ratcliffe (2000) in eliciting preferences for priority in liver transplant.</p> <p>Attributes: Age, alcoholic liver disease (responsibility), expected survival, time spent waiting, re-transplant status</p> |
| Dolan et al. (2008) | <p>Attributes identified via focus groups with 57 public and 172 NHS employees.</p> <p>Attributes: Age, severity, responsibility for illness; added rarity at request of NICE</p> |
| Green & Gerard (2009) | <p>Attributes identified through empirical literature review and discussions with experts and decision makers.</p> <p>Attributes: Severity, health improvement, value for money, other treatments</p> |
| Baker et al. (2010) | <p>Attributes identified through qualitative focus groups and ‘Q-methodology.’</p> <p>Attributes: Age, quality-of-life, length-of-life</p> |
| Desser et al. (2010) | <p>Attributes were specific to objective – is there a preference for prioritising drugs for rare diseases?</p> <p>Attributes: Disease prevalence, relative cost</p> |
| Koopmanschap et al. (2010) | <p>Attributes identified through discussion with “experienced HTA researchers.”</p> <p>Attributes: Budget impact, productivity gains, disease severity, cost-effectiveness, QALY gain per patient, composition of QALY gains (quality vs. survival), uncertainty in ICER</p> |
| Diederich et al. (2012) | <p>Attributes were chosen by investigators to test “whether specific patient groups should receive preferential access to medical services.” Identification process not specified.</p> <p>Attributes: Health status, quality-of-life, healthy lifestyle (responsibility), patient age, carer status, occupational status (SES)</p> |
| Linley & Hughes (2012) | <p>Investigators “reviewed relevant documents and policies to identify nine specific prioritisation criteria (besides clinical-effectiveness and cost-effectiveness)”.</p> <p>Attributes: Severity, unmet need, innovation, societal benefit, disadvantaged population, age, end-of-life, cancer, rare disease</p> |
| Shah et al. (2012) | <p>Attributes were specific to testing “whether the policy of giving higher priority to</p> |

| | |
|----------------------|---|
| | <p>life-extending end of life treatments ... is consistent with the preferences of members of the general public.”</p> <p>Attributes: Life expectancy without treatment, quality without treatment, gain in life expectancy with treatment, gain in quality with treatment</p> |
| Norman et al. (2013) | <p>Attributes selected via literature review, particularly Olsen et al. (2003).</p> <p>Attributes: gender, smoking status, socio-economic status, healthy lifestyle, carer status, gain in life expectancy.</p> |

Most of the elicitations in this table identified their attributes via literature review or expert opinion, but in a few cases, such as Ubel and Loewenstein’s (1996) elicitation of allocative preferences by prognosis, the attributes were dictated by the specific objective of the study. Dolan et al. (2008) and Baker et al. (2010) used focus groups to identify attributes for their elicitations. Focus groups have the notable advantage of allowing for deliberation and reflection among participants about each attribute. Because priority-setting is a social exercise, it is argued that the reasons underpinning this process should be elicited in a social setting (Hasman 2003). However, as with any elicitation of public preferences, there is nothing to ensure that the opinions that emerge from a focus group will be ‘fair.’ Indeed, given relatively small numbers of often unrepresentative participants, and the potential for ‘bandwagon effects’, focus groups may in fact be more likely to produce an aberrant result than less deliberative but more broadly-based approaches (Dolan et al. 2008). Price (2000) also asserts that it is common for members of focus groups to engage in power struggles and strategic behaviours that have little empirical or moral relevance.

None of the studies in Table 3.2 applied an ethical filter as described in this chapter. Therefore, some of the studies included attributes – most notably patient gender, but also personal responsibility and disease rarity – that the empirical ethics review here found to be ethically unjustified. Other studies included attributes such as occupational status or social role, for which the review found little evidence of public support. Given the importance of context in a stated preference elicitation, it is likely that eliciting preferences over different sets of attributes will generate different marginal weights for those attributes. For example, to the extent that an older patient may also be viewed as more responsible for their illness than a younger patient, including or excluding personal responsibility may affect preferences over age. Including

attributes for which there is no prior evidence of public support also means that other attributes that are in fact more relevant might be excluded, given the finite number of factors that can be included in any elicitation, and particularly stated preference elicitation (Froberg & Kane 1989).

3.4 Fair and relevant attributes

The empirical evidence and ethical justifications for each of the attributes discussed in section 3.2 is summarised in Table 3.3:

Table 3.3: Summary of the empirical ethics review

| Attribute | Empirical evidence | Ethical justification(s) |
|--------------------------|--|--|
| Age | <ul style="list-style-type: none"> ✓ Consistent preferences for younger patients, but not necessarily linear ✗ No support for absolute age cut-offs | <ul style="list-style-type: none"> ✓ Maximisation of life expectancy ✓ Maximisation of productivity ✓ 'Fair innings' egalitarianism |
| Social role/productivity | <ul style="list-style-type: none"> ✗ Only limited support for prioritising parents/carers ✗ Very little support for discrimination by productivity | <ul style="list-style-type: none"> ✓ Maximising principle of greatest happiness for greatest number |
| Lifestyle/responsibility | <ul style="list-style-type: none"> ✓ Broad preference for prioritising patients with healthy lifestyle ✗ Minority often <i>strongly</i> opposed to prioritising by lifestyle ✗ How to allow for epidemiological determinants? | <ul style="list-style-type: none"> ✗ 'Healthism' – paternalistic attitude that individuals have moral obligation to society to live healthily |
| Prior healthcare | <ul style="list-style-type: none"> ✗ Mixed evidence of preferences for and against patients against patients who had received previous life-saving care | <ul style="list-style-type: none"> ✓ Egalitarianism – no “second piece of the pie” ✗ Very exclusionary interpretation of egalitarianism ✗ Implies a narrow definition of healthcare |
| Social inequality | <ul style="list-style-type: none"> ✗ Only limited support for prioritising low SES ✗ Preferences appear to be for overall for equality rather than low SES <i>per se</i> | <ul style="list-style-type: none"> ✓ Specific egalitarianism, to extent low SES are disadvantaged in health ✓ General egalitarianism, if health improves overall well-being of low SES |
| Desert/merit | <ul style="list-style-type: none"> ✗ Little support for prioritisation based on past meritorious actions ✓ Some evidence of preferences for 'punishing' illegal behaviour (e.g. drug use) | <ul style="list-style-type: none"> ✗ No clear principle of justice in support of priority based on merit or desert ✓ Some justification where health needs are result of voluntary efforts to improve societal well-being? ✗ Appropriateness of using healthcare system as “omnipotent Supreme Court” dispensing reward/punishment (Olsen et al. 2003)? |
| 'Self' | <ul style="list-style-type: none"> ✗ Very low levels of support for prioritising on basis of gender, race or sexual orientation | <ul style="list-style-type: none"> ✗ No ethical arguments to justify prioritisation based on identity ✗ Preferences likely associated with prejudice or bias |

| | | |
|-------------------------------------|---|--|
| Initial severity | <ul style="list-style-type: none"> ✓ Preferences for health gains to most severe, even when gains were smaller ✓ Strong preferences for life-saving treatments | <ul style="list-style-type: none"> ✓ Need principles ✓ Rawls' Difference principle ✓ Equality of opportunity |
| Final health state | <ul style="list-style-type: none"> ✓ Preferences for final health state rather than absolute gain ✓ Preferences against patients who remain in severe health state | <ul style="list-style-type: none"> ✓ Maximising interpretation of 'equality of opportunity' ✓ Capabilities theory? * Egalitarianism – exacerbates inequalities * Prioritarianism – abandons worst off |
| Treatment effect (absolute benefit) | <ul style="list-style-type: none"> * Preferences for equal opportunity regardless of absolute gain | <ul style="list-style-type: none"> ✓ Maximisation principles * "QALY Trap" – emphasis on absolute gain may discriminate against disabled * Preferences for maximum benefit may exacerbate health inequalities |
| Duration of benefit | <ul style="list-style-type: none"> * Declining marginal value over duration ? Duration a complex function of life expectancy, age, severity and time preferences | <ul style="list-style-type: none"> ✓ Maximisation principles * Preferences for longer duration may exacerbate inequalities in life expectancy |
| Direction of benefit | <ul style="list-style-type: none"> ? Inconclusive evidence of preferences for preventative vs. acute care ? Difficulty in interpreting direction of benefit – is prevention just issue of timing? | <ul style="list-style-type: none"> ✓ Could be consistent with Maximising principles if prevention maximises outcomes* * Implies preference for healthy over ill, violating Rawls' Difference principle and Prioritarianism |
| Distribution of gains | <ul style="list-style-type: none"> ✓ Consistent preferences for smaller gains to many over larger gains to few ✓ Aversion to 'extreme distributions' | <ul style="list-style-type: none"> ✓ Gain egalitarianism ✓ Maintenance of hope * Contrary to outcome egalitarianism? |
| Rarity | <ul style="list-style-type: none"> * Limited evidence of support for prioritising on basis of relative rarity of a condition | <ul style="list-style-type: none"> ✓ Equality of opportunity ✓ Egalitarianism – shows equal respect for all patients * Egalitarianism – shows less concern for patients with more common diseases? |

Attributes shown in **bold** were included in the pilot preference elicitations. ✓ indicates empirical evidence or ethical justification supporting relevance of an attribute; * indicates empirical evidence or ethical justification opposing the relevance of an attribute; ? indicates ambiguous evidence.

Among these attributes, four appeared to have clear evidence of public support and a defensible ethical justification: patient age, severity without/before treatment, final health state with/after treatment, and the distribution of health gains. A fifth, duration of benefit, also appeared to be relevant, notwithstanding some ambiguity over its relative strength. It is worth acknowledging, though, that some measure of duration would most likely have been included in the elicitation regardless of the empirical or ethical evidence in order to facilitate the calculation of QALYs. Cost attributes such as budget impact and incremental cost-effectiveness were specifically excluded from this review as the overall aim

of was to consider how different patient and program characteristics contribute to the societal value of healthcare. In turn, these societal values could be used to weight an outcome measure – for example, an ‘equity-weighted QALY’ – in an economic evaluation. It would be inappropriate to include cost as a factor in the outcome measure as this would double-count costs in the economic evaluation. This exclusion is consistent with other recent elicitation of societal preferences over efficiency and equity in health (Dolan et al. 2008; Baker et al. 2010; Lancsar et al. 2011; Norman et al. 2013).

It is important to acknowledge the subjectivity of this review, both in interpreting the different theories of justice and in judging the consistency of each attribute with these theories. Luck egalitarianism, for example, was rejected as a defensible theory of justice largely on the strength of Anderson’s (1999) argument that as an egalitarian theory it fails to express equal respect and concern for all citizens. However, as Arneson (2000) noted, this theory also has numerous supporters who see it as coherent and defensible. Likewise, the degree to which each attribute was consistent or inconsistent with different theories of justice was a matter of interpretation, and it was necessary to rely on subjective judgement in weighing the ethical arguments for or against each attribute. This means that although empirical ethics may provide a useful framework for arriving at a fair and relevant set of attributes, it should not be viewed as a strictly *objective* means of accomplishing this task. A different reviewer may have arrived at a different set of attributes. Including more than one reviewer, though, and arriving at a consensus, might mitigate some of this subjectivity. Indeed, it is useful to note here that best practice in empirical ethics suggests a multidisciplinary team that can evaluate the quality of the ethical arguments as well as the empirical data (Mertz et al. 2014), although this was not feasible here.

Richardson (2002) acknowledged the subjectivity inherent in empirical ethics and conceded that it will never be able to provide answers to ethical questions which are unambiguously true or immune to criticism. However, he stressed that “an integral part of empirical ethics should be an acceptance of the fact that argument and evidence are fallible and the conclusions are tenuous and more or less strongly supported in some contexts than others.” In this light, any application of empirical ethics can be seen as a balance between a more objective

interpretation of the empirical evidence – which leaves the process open to Hausman’s (2002) charge of moral relativism – and more subjective interpretations of competing theories of justice. The process described in this chapter favoured ethical subjectivity over empirical relativity.

Some subjectivity is consistent with Walzer’s (1983) argument, mentioned earlier, that different principles of justice should govern different aspects of life. A principle that may be appropriate for one aspect – or in this case, attribute – may be inappropriate for another. In this review, for example, deontological theories were rejected as offering little practical guidance to decision makers, even though they may be eminently practical theories of justice for different aspects of life. Similarly, Konow (2003) noted that the idea of ‘fairness’ includes concerns for not only fundamental concepts of equity and justice, but also for some sense of ‘rightness’ in terms of efficiency and need. This suggests that even if it were possible to achieve philosophical agreement on a universal principle of justice, it would not perfectly predict societal preferences as people are motivated by factors outside the scope of such a principle. A fair allocation of resources must reflect fundamental principles of distributive justice, but it must also feel ‘right’ to members of society, even if what feels right may vary between different communities or societies. This vagueness, both in terms of the appropriate principles of justice, and what feels right to society, may limit reproducibility, but as noted above, it can be seen as an essential characteristic of the empirical ethics approach applied here. Attempts to systematize the application of empirical ethics seems more likely to lead to a relativistic emphasis on empirical observation, or a fruitless search for a singular, universal principle of justice, each at the expense of a joint approach. Future research, though, should seek to establish best practices for the application of empirical ethics. An aspect of this could lie in developing methods of collective deliberation over ethical principles and empirical data that could lead to more consistent and stable results without resorting to aggregation and moral relativity.

The empirical evidence for these different attributes was derived from surveys of geographically, culturally and demographically diverse populations, and therefore does not necessarily represent the preferences of any particular community. This, though, can be viewed as a strength rather than a limitation of

the review, as this diversity will tend to support the identification of a broader range of potentially relevant attributes than a survey of just one population. Furthermore, although the direction of preference may vary between populations, the set of relevant attributes is likely to be consistent. For example, some societies may give greater priority to the elderly, while others may give greater priority to the young, but the relevance of age to priority setting would be equally true in both societies. As members of the community still have the opportunity to assign their own weights (including no weight at all) to each of these attributes in subsequent steps of a Communitarian approach, it is the community that defines the importance of each attribute, regardless of the source of these attributes. It is possible, though, that a particular community or society may hold a strong and universal preference for some obscure patient or program characteristic. In such a circumstance, the broad perspective taken here would fail to recognise or incorporate this unique preference.

The attributes identified by this empirical ethics review were largely consistent with the NICE guidance on social value (National Institute for Health and Clinical Excellence 2008). This guidance specifically excludes 'rule-of-rescue' and lifestyle or responsibility issues, and also states that it is not appropriate to consider gender, race or socio-economic status factors in the distribution of healthcare resources, although it is appropriate to consider these factors in the context of reducing health inequalities. The key divergence with these guidelines is in the inclusion of age. The NICE guidelines state that patients should not be denied or have restricted access to treatment on the basis of age alone and exclude any role for age-related preferences. Rawlins (2005), writing on the role of citizen's juries in prioritising health care resource allocation, also suggested that age should not be a factor in societal value considerations. However, the empirical evidence consistently demonstrated public support for age as a factor in priority setting, and it is an important element of fair-innings egalitarianism as well as utilitarian theories of justice.

This empirical ethics review supports the hypothesis that society may be concerned with more than simply maximising aggregate QALYs. Although initial and final health states are related to absolute health gain, preferences for health gains do not appear to be independent of these start and end points.

Likewise, the apparent interaction between quality and duration casts doubt on the presumption of a strictly linear value function. A preference for younger patients is consistent with QALY maximisation, to the extent that younger patients generally have a greater potential for QALY gains, but this preference persisted even when younger and older patients had the same capacity to benefit, suggesting that such a preference reflects more than maximising principles. Most convincingly, there was a clear preference for egalitarianism over maximisation in the allocation of health gains. Together these findings cast doubt on the underlying societal support for the principle of strict QALY maximisation, and particularly its presumption of distributive neutrality in the distribution of health gains. This is not sufficient, however, to demonstrate support for a broader conception of well-being, as most of the empirical studies were based on simple yes/no or ranking questions, with little or no consideration for the strength of these preferences or for the trade-offs between different attributes. For example, Ubel et al. (1998) found a preference for preventative care using a simple ranking exercise, but showed that when strength of preference information was incorporated, this preference was no longer statistically significant. Likewise, Shah (2009) notes that many preference studies focus on a single trade-off and may fail to capture concerns for, or interactions with, other factors. Estimating the relative strength of the equity-efficiency trade-off for the attributes identified here requires a process that forces respondents to make trade-offs between different elements of value. A review of methods for eliciting such preference weights will be presented in the next chapter.

Chapter 4:

Comparative review of stated preference elicitation methods

The empirical ethics review of Chapter 3 suggested that the public's preferences may not be consistent with the principles of strict QALY maximisation, particularly the presumption of distributive neutrality (Nord et al. 1995; Schwappach 2002a; Dolan et al. 2005). Instead, the public appeared willing to forego some potential health gains in order to prioritise younger patients, those in a more severe health state, and those that could be returned to some reasonable final health state. They also appeared to have a preference for how health gains were distributed independent of patient characteristics, generally preferring smaller gains to more people over larger gains to fewer people.

As noted in the previous chapter, these results in themselves are not sufficient to estimate the magnitude of any equity-efficiency trade-off. Estimating the relative strength of preferences, rather than just an ordinal ordering of priorities, requires a process that forces respondents to make trade-offs between different factors while recognising the sacrifices or opportunity costs associated with those trade-offs (Shackley & Ryan 1995). Menzel (1999) also argues that the shift in perspective associated with Communitarianism, from individual well-being to community well-being, has implications for how society's preferences should be measured. Whereas conventional elicitations of individual welfare ask respondents to judge how they would feel about being in a certain condition or health state, elicitations of societal welfare require respondents to consider interpersonal trade-offs and how they would feel about *others* in a particular condition.

This chapter reviews preference elicitation methods that can be used to elicit societal preferences. These methods are based on Lancaster's theory of value and the theory of compensatory decision making, both of which are described in Section 4.1. Section 4.2 outlines a framework for comparing the characteristics and context of different elicitation methods, and the results of this methodological review are presented in section 4.3. Based on these comparisons, section 4.4 discusses the rationale for preferring two particular stated preference methods: discrete choice experiments and constant sum paired comparisons. Finally, section 4.5 reviews recent applications of these two methods in the context of healthcare, including the setting and format of the surveys and their approaches to statistical modelling.

4.1 Measuring preferences and choices

Preference elicitation methods seek to measure the relative impact or importance of different characteristics or attribute levels in a decision (Louviere et al. 2000a; Louviere & Islam 2008). Economics has typically relied on a revealed preferences approach, which infers preferences from actual decisions made under realistic circumstances and binding constraints. In contrast, the defining characteristic of a stated preference elicitation is the hypothetical nature of the task: respondents are asked to make a hypothetical choice between (often hypothetical) scenarios (Hensher et al. 2005; Louviere et al. 2000b). Stated preference approaches fall into two broad categories: choice tasks and matching tasks. Choice tasks ask respondents to choose one or more preferred options from a set of alternatives, while matching tasks ask respondents to provide a number that would make them indifferent in some sense between two or more alternatives (Carson & Louviere 2011).

The primary advantage of the revealed preferences approach is that it avoids the possibility of a 'hypothetical bias,' which suggests that respondents to a stated preference elicitation may be more or less sensitive to aspects of a hypothetical choice than they would be when making an actual choice (Loomis 2011). However, the disadvantage of a revealed preferences approach is that it is often limited to observable markets and historical decisions. The attributes in a

revealed preferences analysis also often move together, making it difficult to evaluate the impact of an independent change in a specific attribute. In contrast, stated preference elicitation methods are based on experimental designs that can be systematically manipulated to test the impact of each attribute over scenarios, or even markets, that do not necessarily exist in the real world (Hensher et al. 2005; Louviere et al. 2000b). Although it is possible that these hypothetical responses would not necessarily translate into actual choices, there is evidence to suggest that techniques such as ‘cheap talk’ and uncertainty coding can reduce the incidence of hypothetical bias in a stated preferences elicitation (List & Gallet 2001; Murphy et al. 2005).

4.1.1 *The theory of value and compensatory decision making*

Stated preference methods stem primarily from psychometrics, which seeks to assign values to subjective psychological concepts such as attitude and preference, but also draw on economic theory, particularly the theory of value and the principle of compensatory decision making (Brazier et al. 1999). Lancaster’s theory of value holds “...that goods possess, or give rise to, multiple characteristics in fixed proportions and that it is these characteristics, not goods themselves, on which the consumer's preferences are exercised” (Lancaster 1966). That is, utility is derived from the characteristics goods possess, rather than from the goods *per se*. Any class of good, therefore, can be defined by its particular combination of characteristics or attributes. Different candy bars, for example, can be described by a set of characteristics that may include sweetness and chewiness. Any one good may be associated with many characteristics, and many goods may produce the same set of characteristics (Louviere et al. 2000b).

The theory of value is the basis of compensatory decision-making, which assumes that in choosing between alternatives, a less preferred level in one attribute can be compensated for by a more preferred level in another attribute (Hogarth & Karelaia 2005; Kjær 2005). Formally, the utility (U) of alternative i to individual n is an additive function of the positive or negative value (v_{in}) associated with the level of each attribute (a_i) and the decision weight associated with that attribute (w_{in}):

$$U_{in} = \sum f(a_i \cdot v_{in} \cdot w_{in}) \quad (4.1)$$

Compensatory decision making is consistent with a rational comprehensive approach to decision making. Decision makers adopting a rational comprehensive strategy are assumed to estimate the expected net utility associated with the attributes and levels of each alternative, and to choose the alternative that maximises expected value (Rosenhead 1980; Wright 1975).

The precise willingness to trade a quantity of one attribute for another is defined by the marginal rate of substitution (MRS):

$$MRS_{a_1, a_2} = \frac{\delta v / \delta a_1}{\delta v / \delta a_2} \quad (4.2)$$

Where MRS is the ratio of the marginal change in the value (v) of a good or alternative given marginal changes in attributes a_1 and a_2 . If the MRS of each attribute characterising a good is calculated relative to the same attribute, the relative importance of each attribute can be expressed in terms of the willingness to trade or sacrifice that common attribute, known as the numeraire. When this numeraire is price or income, MRS can be interpreted as the marginal willingness-to-pay for a marginal change in the level of attribute a_i (Lancsar et al. 2007; Lloyd 2003).

4.1.2 Random utility theory

The conception of utility in a stated preference elicitation is generally based on random utility theory (RUT), which holds that the study of any particular decision process is probabilistic and cannot be perfectly predicted (Kjær 2005; Louviere et al. 2000a). Under RUT, the latent (unobserved) utility (U_i) associated with a particular good or alternative is derived from an observed, systematic component (v_i) and an unobserved component (ε_i):

$$U_i = v_i + \varepsilon_i \quad (4.3)$$

Although the respondent is assumed to be a rational, utility-maximising consumer consistent with classical microeconomic consumer theory, including

complete, stable and consistent preferences, the unobserved component of utility renders any decision stochastic from the perspective of an observer.

Respondent preferences are incorporated into the choice model by specifying the systematic component of utility (v_i) as a function of observed attributes:

$$v_{it} = \sum \beta_k x_{kt} \quad (4.4)$$

where k is the number of observed attributes, β_k is the impact or importance of attribute k on observed utility and x_k is a vector of observed values for attribute k . It is these β 's, or 'part-worth utilities,' that stated preference methods seek to measure, either directly or indirectly (Louviere et al. 2000a). Direct elicitation methods ask respondents to indicate the degree of importance they attach to each attribute, while indirect measures infer attribute importance by analysing repeated choices or matching estimates (Louviere & Islam 2008). Direct approaches can often be associated with strategic behaviours, such as respondents offering 'protest bids' in order to manipulate the results of the elicitation, while indirect approaches are felt to limit the opportunity for such strategic behaviours, in part because respondents may be less likely to recognise the objective of the elicitation (Carson et al. 2001). Perhaps for this reason, Louviere and Islam (2008) found little correlation between preferences elicited using direct and indirect methods.

4.2 A framework for comparing stated preference methods

Shackley and Ryan (1995) argue that any elicitation of stated preferences should measure preferences on a cardinal scale, allow consideration of opportunity cost, and incorporate an appropriate context. First, a cardinal scale allows for the measurement of the distance between alternatives or attribute levels on some interpretable scale of importance (Ali & Ronaldson 2012). An interval cardinal scale is fixed at an arbitrary point and allows measurement of the distance between points in common units (i.e. the distance between 4 and 5 is equal to the distance between 9 and 10), but one cannot say that 10 is twice as much as 5. A ratio cardinal scale, on the other hand, has a natural zero that

allows relative comparisons such as “twice as much” or “half as much.” Economic comparisons require a cardinal scale, but in general an interval scale is sufficient to measure the incremental difference between two alternatives (Brazier et al. 1999). Second, opportunity cost is the explicit recognition of potential benefits that must be foregone as the result of choosing to allocate scarce resources in an alternative way. It is the recognition of such costs that distinguishes *preferences* from *choices*. A car buyer may *prefer* a luxury model, but financial constraints and/or consideration of the opportunity costs may result in the buyer *choosing* a more economical model (Louviere et al. 2000b). Finally, context refers to the combination of elements such as the choice format, the detail provided, and the attributes and levels themselves, all of which interact to form the context of the decision task. For example, a task that provides descriptive text or a photograph of a single alternative has a very different context than a task that describes two or more competing alternatives in quantitative terms. There is no ‘correct’ context, but as a number of authors note, decision makers are used to making decisions within a particular context, and there is substantial evidence to suggest that changing that context to suit a particular elicitation method may adversely impact the face validity, accuracy and predictive ability of the task (Giacomini et al. 2012; Hensher & Collins 2011; Louviere et al. 2000b; Shackley & Ryan 1995). For this reason, stated preference methods should generally be appropriate to the usual context of the decision that is being elicited: the method should adapt to the decision context, not the other way round (Mullen 1999). This simple guidance is complicated here though, by the fact that this is not a ‘usual’ decision – most respondents will have never thought about the degree to which they prefer equity over efficiency in the allocation of healthcare resources, let alone have a usual context for this decision. As such, the appropriate context is not clear. Context is still relevant, though, as Huber (2009) outlines a number of contextual properties of stated preference elicitations that can influence responses in a systematic manner:

- **Comparative vs. individual-alternative orientation:** Comparative tasks tend to put more emphasis on quantitative attributes whose differences are easy to discern or compare across alternatives, while individual-alternative tasks put more emphasis on qualitative attributes that can be interpreted in

the absence of an external reference. Tasks with a comparative orientation tend to encourage respondents to ensure one alternative is 'better' than the other, while tasks with an individual-alternative orientation focus on the overall quality of the alternative. In general, comparative tasks are more contextual and allow for greater consideration of opportunity costs than individual-alternative tasks.

- **Competitive beliefs:** Competitive beliefs, or pre-existing expectations and associations, may be used as heuristics to simplify choice tasks. For example, a consumer may associate high price with high quality, regardless of what is actually shown in the task. Decontextualising a task, as in an indirect elicitation or single-alternative scenario, breaks down these conscious or unconscious associations and forces decision makers to assess the importance of each attribute independent of the others.
- **Reflective vs. immediate:** Reflective tasks tend to emphasise longer-term trade-offs that may be less tangible, while immediate tasks are more competitive and tend to emphasise attributes with more direct and immediate impacts (e.g. price). Matching tasks tend to be more reflective, as respondents must consider the absolute quality of both alternatives, while choice tasks tend to be more immediate, and emphasise finding the 'best' (or avoiding the 'worst') alternative.
- **Attentional shifts:** Simply mentioning an attribute tends to increase its impact, and attributes that would normally have been ignored may now appear important. Attentional shifts may be avoided by increasing the number of attributes in a task, so that unimportant attributes receive less attention, but this risks respondents over-simplifying the task and ignoring *most* of the attributes. Direct elicitations tend to draw a respondent's attention to less important attributes to a greater extent than indirect elicitations.
- **Simplification risk:** Respondents can simplify a decision task across and/or within attributes. In simplifying across attributes respondents may focus on a few important or 'dominant' attributes, while disregarding attributes

deemed less important (Cairns et al. 2002). Within attributes, respondents may dismiss alternatives with low levels on important attributes (i.e. “loss avoidance”). More complex, immediate or competitive tasks tend to increase the risk of simplification, while more reflective tasks may reduce the risk of simplification.

Each of these properties is present to a greater or lesser degree in all stated preference methods, and as such, each method has different strengths and weaknesses. Therefore, to identify preferred methods for the elicitation of societal preferences over efficiency, equity and distributive justice goals in the allocation of healthcare resources, a comparative review of stated preference methods was conducted in terms of the properties outlined above.

4.3 Review of stated preference methods

As noted above, stated preference tasks can be categorised as matching tasks or choice tasks. Open-ended contingent valuation is a common indirect matching task, where respondents are asked to estimate a willingness-to-pay (WTP) that would make them indifferent between obtaining a particular good and keeping the money. Most individuals, though, have difficulty estimating their WTP for a market good, and have even more difficulty estimating their WTP for a non-market good, often leading to missing or inaccurate responses. For this reason, as well as objections – often in the form of protest bids – to valuing health outcomes in terms of money, open-ended contingent valuation is not commonly used in healthcare (Klose 1999). Although there are other forms of matching tasks used in healthcare such as standard gamble and time trade-off tasks, choice-based approaches are felt to present more familiar decision tasks to respondents, and partly for this reason, are more commonly used (Ali & Ronaldson 2012; Brazier et al. 1999; Carson et al. 2001; Smith 2000).

Choice-based approaches measure ‘dominance,’ or whether one alternative is more, less, or equally preferred to another. Strongly ordered measures of dominance allow a complete ranking of all alternatives with no possibility of two alternatives being equally ranked (‘tied’). Weakly ordered measures can identify one or more preferred alternatives from a set of

alternatives, but assume that the remaining alternatives are equally preferred (Louviere et al. 2000a; Louviere et al. 2000b). These measures are ordinal, in that they can establish the ordering of preferences but not the relative strength of preferences. They can, however, be transformed to a cardinal scale by analysing repeated responses to the same comparison, or responses to multiple comparisons by the same respondent (Ali & Ronaldson 2012; Brazier et al. 1999; Ryan et al. 2001).

A review of common direct and indirect choice and matching methods is presented below, with an emphasis on their basis in theory and a discussion of their contextual properties as outlined above. A sample of each task is also shown. These methods include conventional and conjoint ranking tasks, direct and indirect constant sum scaling, full-profile rating, binary and multinomial choice tasks, and person trade-off tasks. The review is summarised in Table 4.1 at the end of this section.

4.3.1 Ranking

Ranking tasks can be indirect, where respondents order a set of alternatives described in terms of their attributes and levels, or direct, where they order specific attributes or levels. These orderings, by ascending or descending importance or desirability, provide a strongly ordered set of preferences as each option can be identified as more preferred or less preferred to every other option, and can be recast as a series of implicit head-to-head choices in order to transform them to a cardinal scale (Ben-Akiva et al. 1991; Brazier et al. 1999). Miethe (1985) found that in terms of test-retest reliability, convergence between scales, and consistency with theoretical predictions, simple ranking tasks outperformed rating scales and magnitude estimation in measuring ordinal values. Whereas the ranking tasks forced differentiation between values, many respondents to the rating and magnitude estimation tasks opted to say that values were equally important, and the resulting lack of variability and differentiation adversely affected the measurement properties of the tasks. This is a common shortcoming of many simple preference surveys. Overall, he concluded that rank ordering had desirable measurement properties in terms of establishing ordinal importance.

However, ranking tasks are cognitively demanding when there are more than a few options. This is especially true of indirect ranking tasks, where respondents are asked to rank a set alternatives, each of which is itself composed of a set of attributes and levels (Flynn et al. 2007; Lee et al. 2007). For this reason, indirect ranking tasks are rare. The primary drawback of any ranking task, though, is that there is no explicit consideration of the opportunity cost of ranking one alternative more highly than another (Ryan et al. 2001). Although rankings can be expanded into a series of head-to-head comparisons, it is not at all clear that this is how respondents interpret the task. Ben-Akiva (1991) also questions how far these comparisons should be extended. He suggests that respondents are likely find it easy to rank their more preferred options, but may be less likely to pay attention when ranking their less preferred alternatives, making these rankings unreliable. Similarly, forced ranking tasks may lead to an arbitrary ranking of elements over which respondents hold no significant preferences, adversely affect the measurement properties of the task (Lee et al. 2007).

Box 4.1: Direct ranking task

Please arrange the following list of attributes in order of importance, from the attribute you consider **most important** in deciding whether to fund this healthcare program, to which attribute you consider **least important**:

| Importance | Attribute |
|------------|--|
| 1 | Average patient will gain 3.0 LYs |
| 2 | Initial utility of patients is 0.2 |
| 3 | Utility after treatment is 0.5 |
| 4 | Utility after treatment is 80% of full potential |
| 5 | 1000 patients can be treated |

The focus of a direct ranking task is on the relative importance of each attribute, which will tend to decontextualise the task and force respondents to consider each attribute in itself, breaking down simplifying associations between attributes. However, as mentioned earlier, relative preferences are likely to depend on the marginal context of the task, which can only be meaningfully understood in the context of the opportunity cost associated with a particular choice. The decontextualised nature of a direct elicitation ranking task will tend

to emphasise the absolute levels of qualitative attributes, which can more easily be interpreted in the absence of a specific comparator, and therefore may emphasise quality improvements over quantitative gains. A ranking task may also increase the attention given to attributes that may not have been considered in an actual decision, while the reflective nature will tend to focus more attention on longer-term benefits and trade-offs, and relatively less on immediate gains. The complexity of a ranking task increases dramatically with more than a few elements, suggesting that simplification risk – perhaps in the form of an arbitrary ranking of less important attributes – may be high.

4.3.2 Conjoint ranking (best-worst scaling)

As in a conventional ranking task, best-worst scaling (BWS) conjoint ranking tasks present respondents with a set of options, but rather than asking them to rank all options, respondents identify only their most preferred and least preferred elements. This is based on the assumptions that respondents can more easily identify the best and worst or most and least important elements in a choice set than rank all elements, and that the probability of choosing a particular best-worst pair is proportional to the distance between them on a latent utility scale (Flynn et al. 2007; Louviere & Islam 2008). The more common ‘single profile’ BWS task, illustrated in Box 4.2, presents a single scenario or profile to respondents and asks them to identify their most and least preferred elements. By using an experimental design to repeat the best-worst ranking task over different subsets of attributes and levels, BWS can establish the rank of each attribute level relative to a single, least-preferred attribute level on a cardinal scale of ‘relative importance’ (Auger et al. 2007; Flynn et al. 2007; Marley & Louviere 2005). In contrast to choice tasks, this allows a cardinal measure of utility relative to a single attribute (i.e. ‘worst’) rather than to an entire alternative or scenario (Fraenkel 2013; Lancsar et al. 2007).

Box 4.2: Single-profile best-worst scaling task

From the following list of attributes, please indicate which one attribute you consider **most important** and which one attribute you consider **least important** in deciding whether or not to fund this healthcare program:

| Most Important | Attribute | Least Important |
|----------------|--|-----------------|
| | Average patient will gain 3.0 LYs | |
| ✓ | Initial utility of patients is 0.2 | |
| | Utility after treatment is 0.5 | |
| | Utility after treatment is 80% of full potential | ✓ |
| | 1000 patients can be treated | |

The less-common ‘multi-profile’ BWS task is similar to a discrete choice tasks in that respondents are asked to choose between entire scenarios, but unlike discrete choice tasks, respondents must identify their least preferred alternative in addition to their most preferred alternative (Flynn 2010). Although this additional step means that more information on the dominance relationships is collected from a multi-profile BWS than a discrete choice task if there are more than two scenarios in the choice set, it also makes the task more difficult for respondents. The appropriate statistical model is also a matter of some debate (Flynn 2010). For these reasons, multi-profile BWS tasks are not common in health economics and the remainder of this section will consider the more established single-profile BWS.

Like conventional rank ordering, single-profile BWS forces differentiation between attributes and has an unambiguous interpretation, as there should be only one way for a respondent to interpret “most important” or “least important” (Lee et al. 2007). However, it has the advantage of doing so in a much less cognitively demanding manner, as because respondents are only presented with a subset of the overall ranking task at any one time, they are typically able to identify the extremes of a choice set more easily than they can rank those attributes somewhere in the middle (Lee et al. 2007; Marley & Louviere 2005). Lee et al. (2007) found that single-profile BWS results were closely correlated with rank ordering, but required much less cognitive effort on the part of respondents. BWS is also more statistically efficient than ‘pick-one’ discrete choice approaches. A 3-item best-worst choice set generates a complete

set of preference orderings, while a 4-item set can identify 9 of the 11 possible dominance relationships (Marley & Louviere 2005). As in a conventional ranking task though, there is no explicit consideration of opportunity cost. BWS results are less strongly ordered than a traditional ranking task, as a BWS task generates an incomplete ranking for choice sets of more than 3 items (Louviere et al. 2000a).

The contextual characteristics of a single-profile BWS are similar to a conventional ranking task. The focus of the task is on the relative importance of each decontextualised attribute, breaking down simplifying associations between attributes and tending to emphasise qualitative over quantitative attributes. The direct nature of the elicitation will tend to increase the attention given to less important attributes, and the reflective nature of the task will tend to focus relatively more attention on less immediate outcomes. Unlike conventional ranking tasks, it imposes relatively few cognitive demands on respondents and may therefore be relatively less likely to encourage simplification or heuristics in identifying best-worst pairs.

4.3.3 Direct constant sum scaling

Direct constant sum scaling (CSS), also known as ‘budget pie’ or ‘allocation of points,’ asks respondents to allocate a fixed number of points or shares between different attributes to indicate their relative degree of importance. CSS is considered a matching task, although as Carson and Louviere (2011) note, it may be seen more intuitively as utility maximisation subject to a budget constraint as it is not necessarily clear what quantity is being matched in the task. There is no specific theoretical basis for CSS, but it has been argued that because the technique forces respondents to consider trade-offs in their allocation of shares within constrained budget, the technique is consistent with economic theory and possesses cardinal, ratio measurement properties, and in this sense, may be theoretically related to contingent valuation approaches (Ryan et al. 2001). Like a BWS task, a direct CSS attempts to identify the relative importance of attributes and levels within an alternative. Attribute importance weights are calculated by dividing the points allocated to each attribute by the total points allocated. Unlike a BWS task, though, a direct CSS task allows

attributes to be valued as equally important. While this may allow for genuine indifference, it also allows respondents to opt-out of difficult trade-offs between attributes (Louviere & Islam 2008). It is also not clear what attribute weights represent: the relative importance of the attribute, the desirability of the attribute level, or some combination of the two (Louviere & Islam 2008).

Box 4.3: Direct constant sum scaling task

Please allocate **100 points** across the attributes listed below in terms of their relative importance to you in deciding whether or not to fund this healthcare program:

| Attribute | Points |
|--|------------|
| Average patient will gain 3.0 LYs | 25 |
| Initial utility of patients is 0.2 | 40 |
| Utility after treatment is 0.5 | 25 |
| Utility after treatment is 80% of full potential | 5 |
| 1000 patients can be treated | 5 |
| Total | 100 |

The decontextualised nature of a direct CSS task will tend to emphasise the relative importance of attributes and levels within an alternative, and not allow respondents to consider the opportunity costs associated with the alternative as a whole. The instruction to allocate points across all attributes is also likely to draw attention to attributes that may have otherwise been unimportant, and this effect may increase with the size of the initial allocation of points. Unlike BWS, direct CSS does not necessarily force respondents to trade-off between attributes, and is not likely to break down a respondent’s pre-existing associations between attributes, as respondents can allocate equal shares to every alternative. In terms of the earlier example, respondents who associate high price and high quality do not need to distinguish which attribute is more important in itself, and as such, direct CSS may actually reinforce pre-existing (but unobserved) associations and confound the measurement of the importance of individual attributes (Huber 2009). Like BWS, the individual-alternative orientation and lack of context in a direct CSS task is likely to emphasise qualitative attributes that can be judged in isolation. This would be reinforced by the reflective nature of the task. Simplification risk seems moderate to high, as the task is more demanding than a conventional ranking task in that it effectively

asks respondents to estimate the cardinal importance of each attribute in addition to the its ordinal rank. These cognitive demands risk respondents essentially opting out of the task by assigning the same number of points to each attribute or alternative.

4.3.4 Indirect constant sum paired comparison

In contrast to direct constant sum scaling, which asks respondents to allocate some fixed quantity between attributes and levels *within* an alternative, indirect constant sum paired comparison (CSPC) asks respondents to allocate a quantity *between* alternatives. This allocation is assumed to reflect the relative importance or priority the respondents attach to each alternative (Mullen 1999). The initial allocation of this quantity, though, is a critical element in the design of the task. Respondents may have difficulty coping with realistic monetary sums outside of their normal experience, but hypothetical points or unrealistic budgets are likely to result in unrealistic responses. As such, it is more common that respondents are asked to allocate budget shares than actual monetary sums (Mullen 1999).

As with the CSS, CSPC is considered a matching task, although it is not necessarily clear what quantity is being matched (Carson & Louviere 2011). Louviere et al. (2000a) suggest that allocation tasks such as CSPC are consistent with RUT if it can assumed that differences in the allocations reflect differences in latent utility between the alternatives. They also show that responses to CSPC tasks can be transformed to dominance rankings on the basis of which alternative was allocated the majority of the budget. These ranking are more weakly ordered than a conventional ranking task owing to the possibility of equal allocations between alternatives, but may be more strongly ordered than discrete choice tasks given the intensity of preference information that can be inferred from the relative allocations.

In the context of healthcare, Schwappach (2003) suggested that the CSPC task is unique in explicitly connecting budget constraints, opportunity costs, health outcomes and patient characteristics. Schwappach and Strasmann (2006) also suggested that it is particularly suited to setting priorities in healthcare given the ability of respondents to avoid extreme distributions by allocating shares to

less preferred groups, consistent with a view of the importance of the ‘maintenance of hope’ in the allocation of healthcare resources (Menzel et al. 1999). Finally, McIntosh (2003), citing Swallow et al. (2001), suggested that for choices that are highly emotive – such as the allocation of healthcare resources – dichotomous choice tasks may leave respondents dissatisfied with the limited information they are allowed to provide. Indeed, Swallow et al. (2001) found that respondents were anxious to provide information on their strength of preference, and suggested that restricting this ability may discourage respondents from participating fully, possibly introducing a sampling bias into discrete choice tasks.

Box 4.4: Indirect constant-sum scaling task

Please allocate **100 points** across the drug programs listed below in terms of the relative share of societal resources you would prefer to see allocated to each drug:

| | |
|--|---|
| <p>Program X Average patient will gain 3.0 LYs Initial utility of patients is 0.2 Utility after treatment is 0.5 Utility after treatment is 80% of full potential 1000 patients can be treated</p> | <p>Points for Program X</p> <div style="border: 1px solid black; width: 40px; height: 40px; margin: 0 auto; text-align: center; line-height: 40px;">40</div> |
| <p>Program Y Average patient will gain 5.0 LYs Initial utility of patients is 0.6 Utility after treatment is 0.9 Utility after treatment is 90% of full potential 500 patients can be treated</p> | <p>Points for Program Y</p> <div style="border: 1px solid black; width: 40px; height: 40px; margin: 0 auto; text-align: center; line-height: 40px;">60</div> |

The indirect, comparative orientation of CSPC provides much more context than the direct CSS approach, as the allocation of points between alternatives forces consideration of the absolute value of both alternatives as well as the opportunity costs associated with funding one alternative over the other. Although there is a competitive aspect to the task that may encourage simplification and emphasise a few quantitative attributes, the need to consider the relative quality of both alternatives in allocating points suggests that the task may be somewhat more reflective than binary or discrete choice tasks (Schwappach & Strasmann 2006). This relatively greater reflection may also encourage consideration of longer-term and qualitative aspects of the

alternatives. The cognitive demands of CSPC are relatively high as respondents are asked to simultaneously judge the direction and the relative magnitude of their preferences, although a review of preference elicitation methods by Ryan et al. (2001) report favourable completion rates in CSPC tasks. Cognitive demands are likely to be lower than in a CSS task, though, as respondents only have to evaluate two alternatives, rather than a potentially much larger set of attributes and levels. These cognitive demands, as well as ethical objections to any priority setting exercise, may lead respondents to allocate points equally between each alternative to simplify the task. A strategy of loss avoidance may also encourage respondents to moderate their allocations out of a desire to avoid committing too heavily to what may turn out to be the 'wrong' alternative (Baron et al. 2001). Also, similar to biases encountered with rating scales, there is potential for extreme response or end-point bias in the allocation of points between alternatives, where respondents may systematically prefer or avoid the extremes of the constrained budget allocations for reasons unrelated to attribute levels or ethical beliefs (Kaplan et al. 1993; Lee et al. 2007). Overall, the context of a CSPC task seems very high, given its simultaneous consideration budget constraints, trade-offs and opportunity costs.

4.3.5 Magnitude Estimation

Magnitude estimation (ME) is an indirect matching task deriving from psychometrics that asks respondents to provide an estimate of how much better one alternative is than another on a ratio scale. These ratios estimates are aggregated across respondents as a geometric mean and the resulting measure is argued to have cardinal, ratio properties (Brazier et al. 1999; Kaplan et al. 1993). However, ME has no clear basis in economic theory and, as Richardson (1994) notes, the interpretation of the ME question, "how many times is x better (or worse) than y is 'deeply obscure.'" In a comparison of ranking, rating and ME approaches, Miethe found that magnitude estimates demonstrated the lowest degree of convergence with the other results (Miethe 1985). The results of an ME elicitation are more weakly ordered than a ranking task given to the possibility of ties in the preference ordering, but more strongly ordered than a

discrete choice approach owing to the additional strength of preference data collected through the ratio estimation task.

Box 4.5: Magnitude estimation task

Please indicate the value of Program Y by giving it a score *relative to the score of Program X*. For example, if you believe Program Y is twice as good as Program X, you should give it a score of 20. If you believe it is half as good, you should give it a score of 5.

| | |
|--|---|
| <p>Program X Average patient will gain 3.0 LYs Initial utility of patients is 0.2 Utility after treatment is 0.5 Relative gain is 75% of potential health 1000 patients can be treated</p> | <p>Program Y Average patient will gain 5.0 LYs Initial utility of patients is 0.6 Utility after treatment is 0.9 Relative gain is 90% of potential health 500 patients can be treated</p> |
|--|---|

Program X = 10

Program Y =

15

The pair-wise format of a ME task provides respondents a high degree of context and may tend to focus attention on the differences in attribute levels between the two alternatives, as well as reduce the importance of external reference points. However, the requirement to express the overall quality of one alternative relative to the other in the ME task should also force respondents to reflect on the overall quality of each alternative. Despite these relative and absolute comparisons, there is no explicit consideration of opportunity cost in a ME task. The ratio scaling task does not require any explicit trade-offs or choices, and in this respect it is strictly a comparative rating task. The pair-wise comparison format of an ME task makes it easy for respondents to identify differences between attribute levels and will tend to emphasise quantitative attributes, but it also allows the easy comparison of attributes that would otherwise have been unimportant. This may lead to an overemphasis of less important attributes in the scaling task. The ME task appears to be cognitively demanding, requiring consideration of both relative differences and absolute levels, suggesting that respondents may choose to opt-out of difficult ME tasks by choosing a magnitude estimate that set the ratio at or close to one. As in the CSPC, a simplifying strategy of loss avoidance may also encourage respondents to moderate their responses (Baron et al. 2001).

4.3.6 Person trade-off

Person trade-off (PTO) is an indirect matching task that asks respondents how many outcomes of type Y they would consider equivalent in terms of (social) value to X outcomes of another kind. The ratio of Y/X represents the social value of outcome Y relative to X, and is consistent with a random utility interpretation. By repeating the task for different alternatives relative to a common comparator (X), the relative value of each alternative can be plotted on a cardinal scale (Green 2001; Nord 1995a). Many authors argue that PTO is particularly suited to considering the trade-offs inherent in allocating societal healthcare resources as PTO judgements go beyond issues of individual utility to include concepts of fairness and equity (Menzel 1999; Nord 1995a; Pinto Prades 1997; Ubel et al. 2000).

PTO has an intuitive appeal and is argued to have cardinal measurement properties. Baron suggests that “PTO is like [standard gamble (SG) and time trade-off (TTO)] because it asks subjects for a number that makes two options equally preferred in hypothetical decision” (Baron et al. 2001). It also has a number of recognised limitations, however, including start point bias, where the equivalence ratio tends to be correlated with the number of patients in the initial state, and ‘ratio inconsistency’ or ‘multiplicative intransitivity,’ where the equivalence ratios of A:B and B:C are not consistent with the equivalence ratio of A:C (Baron et al. 2001; Schwarzinger et al. 2004; Ubel, Loewenstein, et al. 1996). Finally, and perhaps most importantly, respondents often find the task complex, difficult, and even offensive (Green 2001; Nord 1995a). Damschroder et al. (2007) reported that 91 percent of respondents to one PTO elicitation refused to make a trade-off between groups despite clear differences in severity and health gains. Even when respondents understand and are willing to complete the task, Nord (1995a) reported a high degree of random variation in equivalence estimates, suggesting that PTO may be statistically inefficient relative to tasks with less random variation, and that a large and carefully instructed sample may be required to derive reliable preference estimates.

Box 4.6: Person trade-off task

How many persons would have to be treated under Program Y in order for you to be indifferent to funding Program X or Program Y?

| | |
|--|--|
| <p>Program X Average patient will gain 3.0 LYs Initial utility of patients is 0.2 Utility after treatment is 0.5 Relative gain is 75% of potential health</p> | <p>Program Y Average patient will gain 5.0 LYs Initial utility of patients is 0.6 Utility after treatment is 0.9 Relative gain is 90% of potential health</p> |
|--|--|

Program X = 100

| | |
|--------------------|-----|
| Program Y = | 150 |
|--------------------|-----|

PTO tasks are highly contextualised, as similar to CSPC and ME tasks, respondents must consider the overall quality of both alternatives in formulating a person equivalence value. This will tend to make the task less competitive and more reflective than choice tasks, although the comparative nature may tend to emphasise a few quantitative attributes where differences are easier to discern. The opportunity cost of prioritising one group over the other is implicit in the person equivalence value; indeed, the equivalence value defines the opportunity cost of prioritising one group over the other. Although reflective tasks generally have a lower simplification risk than more immediate choice tasks, the simplification risk with PTO seems higher in light of evidence that many respondents appeared to avoid the trade-offs intrinsic to the PTO by offering protest bids of infinity or equal equivalence values (Green 2001). In order to overcome the difficulty of choosing a specific person equivalence values in a PTO task, many investigators use a ‘ping pong’ format to present a series of successively narrower high and low equivalence values to respondents until they converge at an indifference point (Nord 1995a; Rodriguez-Miguez & Pinto-Prades 2002; Damschroder et al. 2005; Baker et al. 2010). This iterative choice format, though, changes the nature of the PTO from a matching task to a series of linked discrete choice tasks.

4.3.7 Full-profile ratings

Full-profile ratings tasks ask respondents to assign a value or rating to an individual alternative defined in terms of its attributes and levels. This rating can

be measured on a numeric scale such as 0 to 10, or a more qualitative scale such as 'relative importance' or 'likelihood of choice.' Full-profile ratings are popular due in large part to the relative ease of the task: they are not cognitively demanding, they can be performed in relatively little time, and they can typically accommodate more attributes than a ranking task (Lee et al. 2007). However, as Louviere et al. (2000a) note, the approach assumes that respondents are able to consistently and reliably estimate their preference for each alternative. They argue that this is a strong assumption in light of common biases associated with ratings scales, including acquiescence bias, where respondents decline to trade-off and value most or all attributes or alternatives as important; extreme response bias, where respondents systematically use only one segment of the rating scale (i.e. moderate responses concentrated around the scale mid-point or extreme responses concentrated at the upper or lower ends of the scale); and, in the opposite direction, a tendency for respondents to want to use each category in a rating scale equally often (Kaplan et al. 1993; Lee et al. 2007). In addition, there is no theoretical basis for interpreting the difference between, for example, a 6 and a 7 and it is therefore not clear that the distance between different points on a rating scale have interval properties (Kaplan et al. 1993; Louviere et al. 2000b).

Although ratings data can be transformed into cardinal utility if it can be assumed that the rating scale accurately represents underlying latent utility, and that a particular rating implies a latent utility between two critical utility thresholds, it is also possible to transform ratings data into weakly ordered ordinal rankings data after allowing for ties. Such a transformation requires much weaker assumptions about the nature of the rating scale and the abilities of the respondents than does treating the scale as representative of latent utility (Louviere et al. 2000b). However, in a comparison of ratings versus rankings and discrete choice tasks, Boyle et al. (2001) found that ordinally transformed ratings could not recover full rankings or 'choose one' discrete choices, primarily due to respondents opting-out of implicit ranking tasks by choosing ties.

Box 4.7: Full-profile rating task

On the scale below, please indicate how likely you would be to recommend Program X for funding:

| |
|--|
| Program X Average patient will gain 3.0 Life years Initial utility of patients is 0.2 Utility after treatment is 0.5 Utility after treatment is 80% of full potential 1000 patients can be treated |
|--|

| | | | | | | | | | | | | |
|------------------------------|---|---|---|---|---|---|---|---|---|---|----|-----------------------------|
| Not at all likely | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Extremely likely |
|------------------------------|---|---|---|---|---|---|---|---|---|---|----|-----------------------------|

Full-profile rating emphasises the absolute quality of an alternative as a whole, rather than the relative importance of attributes. The tasks are intrinsically reflective, as they do not require trade-offs or direct differentiation between attributes or alternatives, although there is evidence that respondents can quickly recognise and adapt to the range of quality between the alternatives, suggesting at least some comparative element to the task (Huber 2009; Kaplan et al. 1993). Kaplan (1993) argues that this property may allow ratings data to be meaningfully analysed using an analysis of variance approach. Unexpectedly, full-profile ratings tasks have been found to focus respondents' attention on a small number of attributes. Huber (2009) notes, "there is no logical reason why ratings-based conjoint should limit attention to a small number of attributes, but that is what happens, study after study." There also tends to be a simplifying emphasis on loss avoidance as respondents penalise alternatives with low levels on key attributes (Huber 2009). However, the individual-alternative orientation – even with the implicit comparative element between alternatives within a larger elicitation – and the abstract nature and weak theoretical basis of ratings scales means the task is extremely decontextualised and does not allow for any consideration of opportunity cost.

4.3.8 Binary choice

A binary choice task can be thought of as a special case of a full-profile ratings task where the rating scale is reduced to 'yes' and 'no,' or 'acceptable' and 'unacceptable.' Such an approach eliminates the scale biases associated with full-profile ratings tasks and provides an unambiguous interpretation of the

response. Binary choice tasks are less cognitively demanding than full-profile ratings tasks as respondents need only answer yes or no as opposed to assigning a rating, and there is a suggestion that this may reduce the incidence of non-compensatory decision making (Lim & Edlin 2009). Binary choice tasks have the advantage of closely approximating the format of the decision task facing health care decision makers, where they most often must judge the acceptability of an individual alternative rather than assign a rating or make a choice between competing alternatives (Tappenden et al. 2007).

Box 4.8: Binary response task

Please indicate if you consider Program X to be acceptable for societal funding:

| |
|--|
| <p>Program X Average patient will gain 3.0 Life years Initial utility of patients is 0.2 Utility after treatment is 0.5 Utility after treatment is 80% of full potential 1000 patients can be treated</p> |
|--|

Acceptable Unacceptable

Preferences for a particular alternative are calculated relative to a defined or undefined status quo. If the status quo is explicitly defined prior to the choice task, it allows some implicit consideration of the opportunity cost associated with *rejecting* the alternative, although there is no consideration of the opportunity cost of *accepting* the alternative. If the status quo is not explicitly defined, it is possible, and even likely, that each respondent will have a different interpretation of the utility implications and opportunity cost associated with rejecting the alternative. As such, it may be difficult to identify the specific attributes and levels associated with the rejection of the alternative (Kjær 2005; Ryan & Skatun 2004).

The results of a binary choice task are more weakly ordered than full-profile ratings transformed into ranks as there is likely to be a greater proportion of ties given the greatly reduced response scale (Louviere et al. 2000b). At the extreme, all alternatives in a binary choice task could be tied as 'acceptable' or as 'not acceptable'. In this case, there is no differentiation between alternatives and

no meaningful preference data is captured. Otherwise, the context of the task is very similar to the full-profile rating task.

4.3.9 Discrete choice experiments

Discrete choice experiments (DCEs) are an indirect, choice-based approach that asks respondents to select their most preferred option from a set of two or more alternatives. Such tasks are similar to decisions respondents face on a daily basis and appear relatively easy for respondents to grasp. For this reason, discrete choice tasks are increasingly preferred over ranking and rating tasks for eliciting preferences in healthcare (de Bekker-Grob, Ryan, et al. 2010; Kjær 2005; Ryan et al. 2001). As a discrete choice task identifies only one preferred alternative per choice set, the results are very weakly ordered and it is necessary to repeat the choice task across a series of alternative pairs (or triplets) to generate a complete ordering of preferences (Louviere et al. 2000b). Although responses to a DCE are strictly ordinal, cardinal preferences can be derived by assuming, based on probabilistic choice theory, that the probability of choosing one alternative over another is proportional to the difference in latent utility between alternatives (Ali & Ronaldson 2012; Kjær 2005).

Box 4.9: Discrete choice task

If you were able to fund *only one* of the two drug programs described below, would you prefer to fund Program X, Program Y or neither drug?

| | |
|---|--|
| Program X Average patient will gain 3.0 LYs Initial utility of patients is 0.2 Utility after treatment is 0.5 Utility after treatment is 80% of full potential 1000 patients can be treated | Program Y Average patient will gain 5.0 LYs Initial utility of patients is 0.6 Utility after treatment is 0.9 Utility after treatment is 90% of full potential 500 patients can be treated |
|---|--|

Prefer to fund Program X
 Prefer to fund Program Y

Discrete choice tasks are conceptually related to binary choice tasks, but the inclusion of two or more mutually exclusive alternatives, rather than an often implicit status quo, makes the task highly contextualised and highlights the opportunity costs associated with choosing one alternative over another. Unlike

matching tasks such as CSPC and PTO, where respondents must judge the relative value of both alternatives, or even more reflective full-profile choice tasks, DCE tasks only require respondents to identify the 'best' alternative. This is likely to shift attention away from the overall quality of an alternative and toward a competitive focus on ensuring that one alternative is better than another. There is a high risk that this competitiveness may lead to simplification and a focus on differences in a few key attributes – particularly on quantitative attributes where differences are easy to discern. This may also manifest itself as 'loss avoidance', where alternatives with low levels on key attributes are quickly dismissed, even where those attributes may have otherwise been unimportant in the decision process (Huber 2009).

Simplification can also lead to non-compensatory decision strategies such as lexicographic or dominant preferences, where respondents do not trade-off between alternatives but rather always choose the alternative with the preferred level of a specific attribute, regardless of the levels of the other attributes (Scott 2002). Such preferences are not irrational, but complicate the analysis of choice as such preferences cannot be expressed in terms of marginal rates of substitution or an additive utility function as no trading takes place, and thus are inconsistent with the theory underlying the stated preferences approach (Lancsar & Louviere 2006; McIntosh & Ryan 2002; Scott 2002).

Table 4.1: Summary of stated preference elicitation methods

| Property | Ranking | Conjoint ranking (BWS) | Constant sum scaling | Constant sum paired comparison | Magnitude estimation | Person trade-off | Full-profile ratings | Binary choice | Discrete choice experiments |
|-----------------------------------|-----------------------------|-----------------------------|-----------------------------|--------------------------------|------------------------------|------------------------------|-----------------------------|-----------------------------|------------------------------|
| Choice or matching | Choice | Choice | Matching | Matching | Matching | Matching | Choice | Choice | Choice |
| Direct or indirect | Both ¹ | Both ¹ | Direct | Indirect | Indirect | Indirect | Indirect | Indirect | Indirect |
| Cardinal scale | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Comparative vs. individual | Individual | Individual | Individual | Comparative | Comparative | Comparative | Individual | Individual | Comparative |
| Competitiveness | Low | Low | Low | Moderate | Moderate | Moderate | Moderate | Moderate | High |
| Immediate vs. reflective | Reflective | Reflective | Reflective | Mixed | Mixed | Mixed | Reflective | Immediate | Immediate |
| Attentional shifts | Absolute qualitative levels | Absolute qualitative levels | Absolute qualitative levels | Relative quantitative levels | Relative quantitative levels | Relative quantitative levels | Absolute qualitative levels | Absolute qualitative levels | Relative quantitative levels |
| Simplification risk | Low | Low | Mod-High | Moderate | Moderate | Moderate | High ² | High | High |
| Marginal context | Low | Low | Low | High ³ | High | High | Low | Moderate ⁴ | High |

¹ Ranking and conjoint ranking tasks can be conducted as direct or indirect elicitations. The characteristics shown are for a direct elicitation.

² Although there is no conceptual reason for full-profile rating to have a high simplification risk, that is what has consistently been observed in empirical studies (Huber 2009).

³ Schwappach argues that CSPC is unique in simultaneously considering budget constraints, opportunity costs, health outcomes and patient characteristics (Schwappach, 2003).

⁴ Reference to an implicit status quo can allow consideration of the opportunity cost of rejecting an alternative, but not the opportunity cost of accepting an alternative.

4.4 Choosing a preferred method

As noted, each stated preference elicitation method has particular strengths and weaknesses and the most appropriate approach depends on the study objectives. Referring back to Chapter 1, the objective here was to identify the relative strength of preferences for different patient and program characteristics, with particular attention to the trade-offs between efficiency and equity or distributive justice in the allocation of scarce healthcare resources. It is also useful to recall the desirable properties of a stated preference elicitation outlined by Shackley and Ryan (1995): preferences should be measured on a cardinal scale, and incorporate the concept of opportunity cost and an appropriate context.

On the basis of these criteria, highly decontextualised tasks such as ranking, conjoint ranking (best-worst scaling), constant sum scaling, full-profile rating and magnitude estimation are immediately excluded as they do not allow for a consideration of opportunity costs. More specifically, they do not allow consideration of preferences at the *margin*: namely, what is one willing to sacrifice to get marginally more efficiency or marginally more equity? In addition, Louviere and Islam (2008) note that responses to indirect tasks tend to give much richer insight into preferences than those to direct tasks, offering further justification for excluding the direct ranking, conjoint ranking and constant sum scaling tasks from consideration.

Binary choice tasks eliminate the scale biases associated with full-profile ratings tasks and provide an unambiguous result that can be interpreted on a cardinal scale. There is also some consideration of opportunity cost through an implicit or explicit consideration of the status quo state. They also resemble the context of many decision tasks in healthcare, where decision makers more often must judge the acceptability of an individual alternative than make a choice between two competing alternatives (Tappenden et al. 2007). Binary choice tasks have been successfully used in the context of healthcare to analyse the preferences of seniors over cataract surgery (Lim & Edlin 2009), and of NICE committee members in recommending healthcare technologies (Tappenden et al.

2007). Notably, Tappenden and colleagues chose a binary task because, as here, it closely reflected the nature of the decision problem faced by respondents. However, the objective of the current study was not to predict funding decisions, but to measure the relative importance of efficiency and different aspects of equity, and the trade-offs between them. As noted in the previous chapter, cost was not included as an attribute in the elicitations as this would, in effect, double count costs in any subsequent economic evaluation. In the absence of cost, a binary choice task would not present decision makers with enough information to make an informed choice, as there would be no opportunity cost associated with accepting a scenario and therefore little reason not to accept every scenario.

With respect to PTO, Nord (1995a), Menzel (1999) and Ubel et al. (2000) argued that it is particularly suited to considering the trade-offs inherent in allocating societal healthcare resources, as PTO judgements extend beyond utility to include considerations of fairness and equity. As noted earlier however, respondents often find the task complex, difficult, and even offensive, and many respondents refused to make trade-offs between groups despite clear differences in severity and health gains. The results of a very small pre-pilot test of this method performed as part of this study are consistent with these findings, as most respondents reported a great deal of difficulty arriving at a specific person equivalence value. Although some investigators have used a ping-pong format to make the task easier, this may negate the reflective nature that Nord (1995a) and others view as an advantage of the method. A number of authors have noted in the context of contingent valuation that dichotomous iterative choice tasks can be associated with a start point bias, as well as a yea-saying bias, where respondents may feel increasing pressure to accept an alternative as the number of iterations grows (Swallow et al. 2001; Chien et al. 2005). The need to fundamentally alter the response format from matching to iterative choice would seem to suggest that although the conceptual basis of PTO is sound, it may be too difficult – cognitively and ethically – for respondents to complete as originally envisioned. This recalls Mullen’s (1999) observation that “theoretical validity does not always coincide with acceptability, people’s comprehension and even people’s value systems.” Indeed, Pinto Prades (1997), in specific

reference to the difficulty respondents had in expressing their preferences with the PTO, quotes Fischhoff et al. (1993): “if subjects cannot use the response mode most convenient to investigators, then investigators must find a response mode that works for subjects.”

Indirect constant sum paired comparison also seems well suited to the elicitation of preferences over the allocation of healthcare resources, given its simultaneous consideration of budget constraints, opportunity costs, health outcomes and patient characteristics (Schwappach 2003). The CSPC allocation task makes it explicit that prioritising one patient group means that the other must necessarily receive lower priority. The task can force a recognition of the same trade-offs as the PTO if the number of patients treated is included as one of the attributes, but it would seem to do so in a more intuitive, less direct, and arguably less discomfoting manner. Although Schwappach and Strasman (2006) reported that 10 percent of respondents to a CSPC elicitation refused to make differential budget allocations, this was well below the 91 percent of respondents who refused to make a trade-off in a PTO reported by Damschroder (2007), and the 32 percent reported by Nord (1995a). These may reflect a refusal to trade-off over what Bartels and Medin (2007) referred to as ‘protected values,’ and Scott (2002) called ‘rights-based’ preferences. Schwappach and Strasmann (2006) argued that the ability to allocate points or budget shares to less preferred groups may allow respondents to avoid compromises over such values and make the task more acceptable to respondents. Indeed, this is consistent with a principle of fairness in the allocation of healthcare highlighted by Giacomini et al. (2012): namely, that everybody should get something and nobody should get nothing. In this sense, although CSPC may not necessarily elicit a *better* answer than PTO, it may be better at eliciting *an* answer, if respondents are more likely to compromise over budget shares than persons. Although the use of CSPC in health economics is not widespread, and its basis in choice theory is less clear than some of the other methods (Ryan et al. 2001), it has been used successfully in a number of elicitations of preferences and values in the allocation of healthcare resources, in addition to the studies noted above (see for example Chan et al. 2006; Linley & Hughes 2012; Ratcliffe 2000).

Finally, discrete choice experiments have the advantage of being relatively easy for respondents to grasp and having a solid basis in probabilistic choice theory (Kjær 2005; Ryan et al. 2001). They also clearly highlight the trade-offs and opportunity costs associated with choosing one alternative over the other. Indeed, the fact that some investigators have suggested reformatting the PTO matching task as a series of linked choices between alternatives suggests that many of the advantages of the PTO can be reproduced with a DCE. They are increasingly being used in health economics for eliciting individual as well as societal preferences, and have been successfully used to elicit societal preferences over the allocation of healthcare resources (de Bekker-Grob, Ryan, et al. 2010).

Relative to CSPC, a DCE task is likely to be more competitive and less reflective, as the emphasis is on picking the best (or avoiding the worst) rather than matching, in some sense, the value of two alternatives. DCE also forces an extreme 'all-or-nothing' distribution that may not be consistent with respondent preferences for the allocation of healthcare resources, particularly when such distributions may involve trade-offs over rights that respondents may feel should not or cannot be compromised in pursuit of other goals. CSPC may be more acceptable to respondents in this context, given its ability to avoid extreme distributions. In allowing respondents to express preferences for specific resource distributions, including equality or maximisation, it may also be a richer source of preference data than the forced-choice task of the DCE. However, CSPC presents a much more challenging task to respondents compared to DCE, and this may lead respondents to choose equal allocations as a way to opt out of difficult allocation tasks, even if they are not truly indifferent to the two alternatives. In light of the theoretical advantages and disadvantages of both methods, it was decided to proceed with both approaches in a pilot study to compare the response characteristics of DCE and CSPC. From this comparison, discussed in the next chapter, a preferred method would be chosen for the primary elicitation.

4.5 Other studies using CSPC or DCE methods

A methodological summary of other studies using CSPC or DCE stated preference elicitation in a societal healthcare context is presented in Table 4.2 below. It highlights the study sample, administration format, and analysis methods, including the regression model, if applicable, and other comparisons or descriptive statistics reported as part of the study.

Table 4.2: Summary of recent DCE and CSPC methods

| Study | Sample & administration | Analysis methods |
|--|--|--|
| CSPC | | |
| Ubel & Loewenstein (1996) | Prospective jurors (N=169) Self-administered paper survey | Categorical: proportion of respondents by categorical distribution of livers and proportion of respondents by reason for allocation. Qualitative descriptions and quantitative summary of reasons for allocation. |
| Abellan-Perpinan & Pinto-Prades (1999) | Undergraduate students (N=149) Self-administered paper survey | Categorical: proportion of respondents by categorical allocation of monetary budget. |
| Ratcliffe (2000) | University employees (N=303) Self-administered paper survey | Additive random and fixed effects linear models and fixed-effects double-bounded tobit model. Respondent ranking of importance of individual attributes. Proportions by difficulty rating and with dominant or strictly egalitarian preferences. |
| Schwappach (2003) | Undergraduate students (N=154) Self-administered internet survey | Additive double-bounded random effects tobit or random effects linear model. Proportions by difficulty rating and dominant or strictly egalitarian preferences. |
| Chan (2006) | Random households (N=281) Face-to-face interviews | Additive random effects linear model. Respondent ranking of importance of individual attributes. Proportions with dominant or strictly egalitarian preferences. |
| Desser et al. (2010) | Random sample of online survey panel (N=1547) Self-administered internet survey | Categorical: proportion of respondents favouring rare or common disease, or indifferent. Likert scale attitudinal questions |
| Linley & Hughes (2012) | Representative UK online survey panel (N=4118) Self-administered | Categorical: proportion of respondents favouring one group or the other, by each attribute independently. Logistic regression to test association between respondent characteristics and preference across |

| | | |
|----------------------------|--|---|
| | internet survey | each attribute independently. |
| DCE | | |
| Bryan et al. (2002) | Random households (N=909) Face-to-face interviews | Additive random effects binary probit. Proportions with dominant preferences or choosing non-dominant alternative in test of non-satiation. |
| Baltussen et al. (2006) | Convenience sample of decision makers (N=30) Group self-administered paper survey | Additive random effects logistic model. |
| Dolan et al. (2008) | Random households (N=559) Face-to-face interviews | Social welfare function to estimate inequality aversion and marginal rates of substitution for different attribute combinations. Subgroup analysis of preferences by observed respondent characteristics. |
| Green & Gerard (2009) | Random-location quota sampling (N=259) Face-to-face interviews | Additive fixed effects conditional logit model. Proportions choosing non-dominant alternative in test of non-satiation and rating task as difficult. |
| Koopmanschap et al. (2010) | Convenience sample of policy-makers, HTA practitioners and health economics students (N=66) Face-to-face interviews | Pooled additive multinomial logit model. Subgroup analysis of preferences by interacting attributes and subgroup. |
| Diederich et al. (2012) | Representative German sample (N=2031) Computer-assisted personal interviews (CAPI) | Pooled additive multinomial logit model. Attribute relative importance. |
| Lancsar et al. (2011) | Representative UK sample (N=587) Computer-assisted personal interviews (CAPI) | Log-linear and 'powered' log-linear conditional logit model, allowing for clustering of standard errors. Distributional QALY weights based on compensating variations |
| Norman et al. (2013) | Representative Australian sample (N=616) Self-administered internet survey | Additive random effects probit with interactions between categorical main effects and LE gain. Equity weights as ratio of expected utility relative to reference scenario. Subgroup analysis of preferences by observed respondent characteristics. |
| Shah et al. (2012) | Representative UK sample (N=4008) Self-administered | Additive conditional logit model with interactions. Subgroup analysis of preferences by observed |

| | | |
|--|-----------------|--|
| | internet survey | respondent characteristics. Proportions choosing non-dominant alternative in test of non-satiation. |
|--|-----------------|--|

Four of the seven CSPC elicitations took a categorical approach, describing the proportion of respondents that favoured one group or the other, or were indifferent between the two, in their allocations. Dessler et al. (2010), for example, described the proportions of respondents that favoured prioritising patients with a rare disease, patients with a common disease, or were indifferent between them (i.e. an equal allocation to both). The limitation of a categorical approach, though, is that in simplifying the continuous allocations to discrete categories, it discards information that would provide a more nuanced understanding of the relationship between attribute levels and preferences, as well as improve statistical efficiency. In addition, a categorical approach cannot interpret the effect of multiple attributes simultaneously. The other three CSPC studies took a regression approach, relating differences in the budget allocations to differences between attribute levels. Reflecting the bounded nature of the response variable, Ratcliffe (2000) and Schwappach (2003) both tested a double-bounded tobit model but ultimately settled on a random effects linear model as they found only minimal evidence of censoring in responses. Chan (2006) used a random effects linear model without testing the fit of a tobit model. In the analysis of DCE responses, all the studies used non-linear logit or probit models and most adopted a random effects specification to account for the panel nature of the data as each individual contributed multiple choice responses.

As outlined earlier in equations 4.3 and 4.4, random utility theory assumes that for individual i , the latent utility associated with task t is a combination of a systematic component based on the sum of k observed attribute levels (x_{kt}) and their part-worth utilities (β_k), and a random component, ε_i :

$$U_{it} = \sum \beta_k x_{kt} + \varepsilon_i \quad (4.5)$$

A random effects specification further assumes that the random component can be disaggregated into an individual-specific term (μ_i), and a stochastic term (ε_{it}) (Baltagi 2008; Croissant & Millo 2008):

$$U_{it} = \sum \beta_k x_{kt} + \mu_i + \varepsilon_{it} \quad (4.6)$$

The individual-specific term is fixed for all choices by individual i but varies between individuals. This allows for correlation between choices by the same respondent and for heterogeneity between different respondents. In a random effects specification, though, the variability in latent utility between individuals reflects the pre-specified distributions of the individual and stochastic error terms and is not directly linked to heterogeneity in tastes or preferences (Morey & Greer Rossmann 2003).

To identify differences in preferences by observed respondent characteristics, Dolan et al. (2008), Koopmanschap et al. (2010), Shah et al. (2012), Linley & Hughes (2012), and Norman et al. (2013) compared preferences between subgroups of respondents stratified by characteristics such as gender, age, children, health status, employment status, or professional role. The studies found some evidence of heterogeneity in preferences over observed respondent characteristics, but this approach has the disadvantage of being strictly deterministic – all respondents in a particular group (e.g. males, or non-smokers) are assumed to share the same preferences (Boxall & Adamowicz 2002; Morey & Greer Rossmann 2003).

Lesson from these DCE and CSPP analyses will be used to inform the analysis of the pilot elicitation, to be discussed in the next chapter, and the subsequent primary elicitation. In addition, the potential benefits of a latent class approach as an alternative to the random effects specification, as well as to an assumption of strictly deterministic preferences by respondent subgroup, will be discussed in Chapter 8.

Chapter 5:

Pilot survey methods & results

The review of stated preference elicitation methods identified the DCE and CSPC formats as having notable advantages over the others in eliciting societal preferences in a healthcare context. As there was no clear theoretical basis for preferring one method over the other, however, it was decided to conduct an empirical comparison to identify a preferred method for the primary elicitation of societal preferences, as well as to refine the wording and presentation of the choice tasks.

As noted in the previous chapter, DCE and CSPC methods are both consistent with random utility theory, and section 5.1 details this theoretical basis. The remainder of the chapter describes the methods and results of the pilot survey. The methods, outlined in section 5.2, are structured around a process described by Ryan (1999) that has become a standard in designing health economic stated preference surveys. The empirical ethics review discussed in Chapter 3 represented the first stage, the identification of attributes. The second stage was to assign levels to these attributes that were both realistic but that also allowed consideration of the full range of values that may be relevant to respondents. The third stage was the experimental design – the systematic combination of attributes and levels that was presented to respondents in order to observe their choices. This stage had to balance statistical efficiency with ‘respondent efficiency’ (Severin 2001), in the sense that there is a limit to the cognitive capacity of any respondent to process the information presented by an experimental design (Amaya-Amaya et al. 2008). This section also discusses the assumption of rationality that underlies all stated preference approaches, and the tests of rationality that are often incorporated into experimental design. The

fourth stage was data collection, which was conducted with a convenience sample to compare the response behaviours of the two formats and to pilot test the wording and presentation of the choice tasks. The fifth stage of the methods describes statistical methods used in analysing and interpreting the DCE and CSPC choice data. The emphasis was on comparative measures such as response behaviour and ease of completion, but the importance weights derived from the choice data are also described. Finally, the results are presented in section 5.3, and the implications of these results for the identification of a preferred format are discussed in section 5.4.

5.1 DCE and CSPC in the context of random utility theory

DCE tasks ask respondents to choose between two alternatives in straightforward manner, and clearly highlight the trade-offs and opportunity costs associated with choosing one alternative over the other (Kjær 2005; Ryan et al. 2001). In a random utility model of discrete choice, the probability of choosing alternative i from choice set $[i,j]$ is assumed to be proportional to the difference in latent utility (U) between the alternatives:

$$\text{Prob}(i|i,j) = \text{Prob}(U_i > U_j) \quad (5.1)$$

This can be re-written to incorporate the systematic (v) and stochastic (ε) components of random utility for each alternative, consistent with random utility:

$$\begin{aligned} \text{Prob}(i|i,j) &= \text{Prob}[(v_i + \varepsilon_i) > (v_j + \varepsilon_j)] \\ &= \text{Prob}[(v_i - v_j) > (\varepsilon_j - \varepsilon_i)] \quad \forall i \neq j \end{aligned} \quad (5.2)$$

In this model, the probability of choosing alternative i from choice set $[i,j]$ is proportional to the difference in systematic utility. The greater v_i relative to v_j , the greater the probability of a decision maker choosing alternative i . Relating differences in observed utility ($v_i - v_j$) to the observed probability of choice means that systematic utility can be measured on the same cardinal scale as probability (Kjær 2005; Green & Gerard 2009; Ali & Ronaldson 2012).

CSPC tasks present two or more alternatives in the same form as a DCE, but ask respondents to allocate points or shares between alternatives, where the relative allocation is assumed to reflect the relative importance or priority the respondents attach to each alternative (Mullen 1999). Although this is a more cognitively demanding task than DCE, Schwappach (2003) argues that CSPC is unique among stated preference methods in explicitly linking budget constraints, opportunity costs, health outcomes and patient characteristics in the consideration of preferences. The theoretical basis for CSPC is less clear than for DCE, but Carson and Louviere (2011) suggest that CSPC can be seen as utility maximisation subject to a budget constraint. In the context of random utility theory, this implies that the goal of the respondent is to maximise utility (U) by allocating a fixed budget (B) between alternatives i and j :

$$U = b_i(v_i + \varepsilon_i) + b_j(v_j + \varepsilon_j), \sum_{i=1}^j b_i = B \quad (5.3)$$

Where v and ε are the systematic and stochastic components of latent utility as in 5.2 above, and b_i and b_j are the shares of the budget allocated to alternative i and j , respectively, subject to the constraint that these shares must sum to the fixed budget. Louviere et al. (2000a) suggest that the difference in the budget shares reflect the differences in latent utility between the alternatives:

$$(b_i - b_j) \cong [(v_i - v_j) + (\varepsilon_i + \varepsilon_j)] \quad (5.4)$$

Analogous to the probabilistic model of discrete choice shown in 5.2, a budget difference of zero (an equal 50%-50% allocation) implies that there is no difference in the latent utility of the two alternatives, while a positive (negative) budget differences implies that the latent utility associated with alternative i is greater (less) than alternative j . As these differences provide cardinal strength of preference information, CSPC tasks, and ordered-response tasks more generally, tend to produce more strongly ordered preference data than DCE, giving them a potential advantage in terms of statistical efficiency (Louviere et al. 2000a; Swallow et al. 2001). This statistical advantage may be offset, though, if a substantial proportion of respondents find the CSPC too cognitively demanding and adopt simplifying strategies in their responses.

5.2 Stated preference design

This section describes the five stage process that was followed in design, administering and analysing the DCE and CSPC questionnaires: the identification of attributes, the assignment of levels, the development of the experimental design, data collection, and data analysis (Ryan 1999). Each of these stages is described in turn below.

5.2.1 Identification of attributes

The number of unique scenarios possible for a set of attributes is given by L^A , where L is the number of levels within an attribute and A is the number of attributes, and shows that the number of possible choice scenarios increases exponentially with the number of attributes (Hensher et al. 2005). As the statistical power of any stated preference elicitation is a function of the number of respondents and the number of choice tasks completed by each respondent, this means that for each additional attribute in an elicitation, a greater number of scenarios must be presented to respondents to achieve a given statistical power (Orme 2006b). Given finite limits to the number of potential respondents, and the time they are willing to devote to completing an elicitation, this means that there is a practical limit to the number of attributes that can reasonably be included in any stated preference task.

There may also be cognitive limits to the ability of respondents to process choice tasks, limiting the number of attributes that can be included in an elicitation. Louviere et al. (2000b), though, argue against such a theoretical limit, and this appears to be supported in part by empirical work from Weiss (1982). She tested the impact of increasing decision complexity in a choice task in terms of the quantity of information presented to decision makers and found that the marginal uptake as new information was added to a scenario was positive (decision makers used more information as more was presented). However, she also found an increase in cognitive strain and a decline in the proportion of all available information used by respondents as complexity increased (decision makers ignored more information as complexity increased). In a similar study, Wright (1975) tested for a tendency toward non-compensatory decision making as complexity increased, and found that decision makers “become increasingly

unidimensional under moderate information load.” This suggests that decision makers stop considering compensatory trade-offs between attributes and focus on maximising fewer and fewer attributes as decision complexity increases. Both of these results are consistent with Simon’s (1955) model of satisficing in response to decision complexity, where he proposed that decision makers make increasing use of simplifying decision rules as complexity increases, sacrificing utility maximisation to minimise cognitive effort.

A psychological explanation for simplification in the face of decision complexity is provided by Miller (1956), who suggested that humans can only process “seven, plus or minus two” separate pieces of information any one time. This finding is the basis for Froberg and Kane (1989) recommending that no more than nine attributes, and preferable fewer, should be included in a stated preference choice task. DeShazo and Fermo (2002) provide more empirical support in reporting that an increase in the number of attributes in a choice task from between four and seven to nine increased the variance in the random component of utility, and that this variance outweighed any potential increase in decision consistency as a result of a more complete description of the alternatives. On these bases, seven attributes was taken as the maximum number that could be feasibly included in the pilot elicitations. This limit is also consistent with several recent reviews of conjoint surveys in healthcare, which found that most elicitations included no more than 6 attributes (Green & Gerard 2009; de Bekker-Grob, Ryan, et al. 2010; Marshall et al. 2010).

Beyond how many attributes can be included in an elicitation is the critical question of *which* attributes should be included (Hall et al. 2004), but there is little consensus on the most appropriate methods or sources for identifying such attributes – theory, existing measures and scales, literature reviews, focus groups, clinical trials, key informant interviews and expert opinion all can and have been used (Kjær 2005; Coast & Horrocks 2007). In this case, the empirical ethics literature review described in Chapter 3 found four attributes that had empirical evidence of support and a defensible ethical justification: patient age, initial severity, final health state, duration of benefit, and the distribution of health benefits. To allow for conceptions of severity based on health state as well as proximity to death, this concept was decomposed

into two separate attributes: initial health state and life expectancy without treatment. Life years gained with treatment was included to allow for an estimate of the duration of benefit, as well as to test societal support for the principles of strict QALY maximisation, despite ambiguous evidence for duration or absolute gain as relevant factors in the empirical ethics review. To consider distributional preferences, the number of patients that could be treated under each alternative was included as an attribute. Together, these attributes allowed for the calculation of the aggregate QALYs gained with each alternative.¹⁰ Aggregate QALYs gained was fixed in each DCE alternative, but varied with the number of patients treated in each CSPC alternative. As noted previously, cost was not included as an attribute.

5.2.2 Assigning levels

There are no clear rules for assigning numeric or qualitative levels to attributes, but in general the levels should be plausible and realistic to respondents and constructed so that they are willing to make trade-offs between attributes (Ryan 1999). The range between the highest and lowest levels of an attribute should also be large enough to include all relevant levels, but not so large as to be unrealistic. However, as the ISPOR Conjoint Analysis Best Practices Task Force (Bridges et al. 2010) notes, “attribute levels should encompass the range that may be salient to subjects, even if those levels are hypothetical or not feasible given current technology.” The importance of an appropriate range is highlighted by Kjær (2005), who emphasised that “an insignificant coefficient does not necessarily mean that the attribute is unimportant to respondents; the correct interpretation is that the attribute did not influence the choice for given levels.”

A change in the level of an attribute is associated with a change in the utility of that attribute. Increasing the number of levels in an attribute provides more information on the form of the utility function – two levels allow the estimation of a strictly linear utility function, while more levels provide more

¹⁰ $\text{Aggregate QALYs gained} = [(\text{life expectancy} + \text{life years gained}) \times \text{final utility} - (\text{life expectancy} \times \text{initial utility})] \times \text{patients treated}$

information on the shape of the utility function (Hensher et al. 2005) – but increasing the number of levels within an attribute has also been shown to be associated with ‘level bias,’ where increasing the number of levels in an attribute tends to increase the significance of that attribute in respondents’ choices. That is, an attribute with 5 levels will often be more significant than an attribute with 3 levels, even if the end-points are the same (Kjær 2005). In addition, as the number of levels in any attribute increase, so does the number of choice scenarios required to achieve a given level of statistical power. To balance the issues of statistical efficiency and level bias with information on the shape of the utility function, each attribute was therefore assigned three levels. As the objective of the stated preference elicitations was to elicit respondents’ preferences over a wide spectrum of hypothetical program alternatives, the levels of each attribute were evenly spaced and set as widely as possible across a plausible range. The specific levels assigned to each attribute are shown in Table 5.1:

Table 5.1: Pilot survey attributes and levels

| Level | Age | Initial utility | Initial life expectancy | Final utility | Gain in life expectancy | Patients treated |
|-------|-----|-----------------|-------------------------|---------------|-------------------------|------------------|
| 1 | 10 | .1 | 1m | .1 | 1y | 500 |
| 2 | 40 | .5 | 5y | .5 | 5y | 2,000 |
| 3 | 70 | .9 | 10y | .9 | 10y | 5,000 |

The levels for age were intended to test preferences for the young, middle-aged and elderly. Similarly, levels for initial and final health states were intended to test preferences for poor, moderate and excellent health. To simplify the presentation of health-related for respondents, each health state was described on a hypothetical 0 to 10 numerical scale, with 0 representing dead and 10 representing perfect health. The characteristics of the levels presented in the tasks were described using health state profiles based on EQ-5D dimensions, similar to the approach used by Schwappach (2003). A minimum life expectancy before treatment of 1 month was intended to represent imminent death while avoiding implausible combinations associated with zero life expectancy but positive utility. The minimum gain in life expectancy after treatment was chosen to be a minimal yet meaningful gain, while the maximum gain in life expectancy after treatment was chosen to be plausible when

considered in combination with maximum age and initial life expectancy. Defining the appropriate levels for the number of patients treated was analogous to the challenge of defining the budget in the CSPC – respondents would likely have difficulty coping with counts that reflected national populations, while small patient counts risked respondents not recognising trade-offs between levels and effectively ignoring the attribute. As such, an upper level of 5000 patients was chosen to represent a comprehensible number of patients, and the lower level was defined to allow for a meaningful distinction between the levels. The middle level was simply the approximate midpoint. See Appendix 5.1 for the attribute descriptions provided to respondents.

5.2.3 Experimental design

The systematic plan for the presentation of different attributes and attribute levels in order to observe respondent choices is known as the experimental design (Louviere et al. 2000b; Hensher et al. 2005). The most comprehensive experimental design is a full factorial, in which every possible combination of attributes and levels is presented. The key advantage of a full factorial design is that each attribute and level appears an equal number of times and each attribute-level combination appears with every other attribute-level combination at least once. This allows the effect of each attribute-level combination, including two-way and higher order interactions, on choice to be estimated independently of each other, known as orthogonality. However, given the 6 attributes noted above, each with 3 levels, the number of possible combinations in a full factorial design is $3^6 = 729$. This is clearly too many tasks to present to any respondent, and it highlights the key drawback of a full factorial design; namely, that it often results in an impractical and unmanageable experimental design (Louviere et al. 2000b).

A more practical alternative to a full factorial is a fractional factorial design, which presents only a subset of possible combinations to any one respondent. *Orthogonal* fractional designs focus on creating statistically independent designs with no correlations between attributes while largely disregarding statistical efficiency. *Optimal* fractional factorial designs, on the other hand, focus primarily on maximising statistical efficiency – extracting the

maximum amount of information from respondents, subject to constraints such as the number of attributes in the design, the number of levels, and the number of tasks in the elicitation (Carlsson & Martinsson 2003). Recall that a full factorial design includes all possible combinations of attributes and levels, and allows for every main effect, as well as all two-way and higher order interactions, to be estimated independently (Kuhfeld et al. 1994). If not all of these effects are of interest, the size of an experimental design can be reduced without sacrificing precision in the relevant parameter estimates by allowing some correlation between irrelevant parameters.

An efficient optimal fractional factorial design maximises the precision of the parameter estimates – or equivalently, minimise the variance of those estimates – for a given set of constraints. The statistical efficiency of different designs can be compared in terms of A-efficiency, G-efficiency or D-efficiency. All three measures are highly correlated, but D-efficiency is used most often, mainly because it is less computationally burdensome than the other measures (Carlsson & Martinsson 2003; Kuhfeld et al. 1994). A D-efficient design relates to the covariance matrix (Ω) of the model to be estimated (Hensher et al. 2005; Carlsson & Martinsson 2003):

$$\Omega = \sum_{i=1}^I \sum_{t=1}^T \sum_{j=1}^J x'_{itj} P_{itj} x_{itj} \quad (5.5)$$

Where x_{itj} is a vector of attribute levels presented to individual i in alternative j or task set t , and P_{itj} is the probability of choosing that alternative, which McFadden (1974) showed is given by:

$$\Pr(j)_{it} | \beta_i = \frac{e^{\beta_i x_{itj}}}{\sum_{j=1}^J e^{\beta_i x_{itj}}} \quad (5.6)$$

Where β_i is the vector of utility weights associated with alternative x_{itj} . A D-efficient design seeks to minimise D-error, calculated as the determinant of the geometric mean of the inverse of the covariance matrix, Ω :

$$\text{D-error} = [\det(\Omega)^{-1}]^{1/k} \quad (5.7)$$

Where k is the number of parameters to be estimated from the design. Minimising D-error has the effect of minimising the variance-covariance matrix of the model, known as the Fisher information matrix, and maximising the

statistical efficiency of the experimental design (Hensher et al. 2005; Carlsson & Martinsson 2003; Kuhfeld et al. 1994). However, as shown in equations 5.5 and 5.6, the D-error of a particular design depends on the choice probability of each alternative in that design. This leads to the paradoxical result that an efficient experimental design requires prior knowledge about the very parameters that the stated preference elicitation is trying to estimate.

The importance of knowing the choice probability of each alternative stems from the criteria for an optimally efficient non-linear choice design identified by Huber and Zwerina (1996): level balance, orthogonality, minimal overlap and utility balance. The design is non-linear because the response variable, choice, is discrete rather than continuous. Although the response variable in the CSPC tasks is continuous, the CSPC questionnaire was based on the same experimental design as the DCE as the design principles are similar for the two formats. Level balance implies that each level appears with equal frequency in the overall design. Orthogonality refers to the statistical independence of the attributes. Minimal overlap means that the same attribute level should not appear in more than one alternative in a particular choice task. Finally, utility balance means that the probability of each alternative being chosen is roughly equal, and that there are no clearly dominated alternatives in the choice set. Little preference information is generated if one alternative is dominated by the other; selection of the dominant alternative simply demonstrates that a respondent is rational by the axioms of choice theory (Johnson et al. 2007). There is a limit to the desirability of utility balance, though, as at the extreme a perfectly balanced scenario would, in effect, be a random choice between two equally attractive alternatives and would not generate any useful choice information (Kanninen 2002). Optimising utility balance requires information on the choice probability of each alternative, although Carlsson and Martinsson (2003) show that when no prior information is available, it can be assumed that the choice probabilities of all the alternatives are equal, even though this limits the potential efficiency of a design. They argue that the utility balance requirement of an efficient design should be seen as an imperative for pilot work that can inform the design of the primary elicitation.

To ensure that correlations between the effects of interest are minimised, a model must be pre-specified at the design stage. The degrees of freedom required to estimate this model determine the minimum number of choice sets that must be included in the experimental design (Hensher et al. 2005). For a main effects model with categorical parameters, the degrees of freedom are given by $A(L-1)$, where A is the number of parameters to be estimated and L is the number of levels. For continuous parameters, the degrees of freedom required are simply A . To estimate two-way categorical interactions, the additional degrees of freedom required are given by $(L-1) \times (L-1)$, while continuous interactions require 1 degree of freedom each. One degree of freedom is also required to estimate the model. For simplicity in the pilot phase, only the main effects for the six 3-level attributes noted in section 5.2.2 were estimated. As this estimation required 12 degrees of freedom, plus one degree for estimation, the pilot survey required a minimum of 13 choice sets. However, as the SAS[®] design macros showed that a design of this size would not achieve level balance, an optimal design could not be generated for 13 choice sets. Instead, the smallest feasible design with at least 13 degrees of freedom was 18 choice sets.

The design process started with a 3^6 full factorial candidate design with attributes for age, initial health state, initial life expectancy, final health state, life years gained and the number of patients. As a product of the other attributes, aggregate QALYs were not included as a separate attribute in the experimental design. Illogical attribute combinations where the net QALY gain with treatment was negative were excluded from the final design, but combinations where the aggregate QALYs gained was zero were included if an increase in quality was offset by a deterioration in life expectancy, or *vice versa*. Scenarios where health state and life expectancy were unchanged before and after treatment were also excluded. Although it could be argued that such a scenario might represent the maintenance of current health through preventative care, it leads to a confusing choice task. This exclusion can also be justified on the grounds that preferences for the direction of health benefit were considered and rejected in the empirical ethics review. Of the 729 scenarios in the full factorial design, 135 (19%) were excluded as illogical. Note that such exclusions are

likely to introduce some correlations between the attributes in the remaining scenarios (Bridges et al. 2010).

A D-efficient fractional factorial design was generated using Kuhfeld's (2010) SAS[®] macros. The D-error of different combinations of the 594 logical scenarios from the candidate design was evaluated using a modified Fedorov algorithm (Johnson et al. 2007; Kuhfeld 2010). This algorithm generated a random design from the candidate design, subject to the specified number of choice sets and alternatives per choice set. For each choice set in this initial design, the algorithm replaced one alternative with a random alternative from the eligible scenarios and evaluated the change in D-error. If it was an improvement, the algorithm moved on to the next choice set. This was repeated for all choice sets until D-error was minimised for that particular design. This process was repeated 100 times and the design with the lowest D-error was selected. As Kuhfeld et al. (1994) note, this process will find an efficient design, but there is no guarantee that it will find the *most* efficient design. Also, as noted above, this algorithm is most effective when prior preference weights can be incorporated into the design algorithm (Carlsson & Martinsson 2003), but as these weights were not known in the pilot phase the algorithm assumed that all scenarios had equal choice probabilities.

The final design had 18 choice sets, but as the literature suggested that this was likely an excessive number of choice sets to present to a single respondent (Bridges et al. 2010), it was evenly divided into two subsets, or blocks, of 9 choice tasks per respondent. The design macros optimised the blocking strategy so as to avoid any interactions with the blocking variable itself. Block 1 of the design was used for the DCE questionnaire, and block 2 was used for the CSPC questionnaire. Although this simplified questionnaire administration, it violated the principles of optimal experimental design and further limited the statistical efficiency of the elicitations (Carlsson & Martinsson 2003). For this reason, the preference data derived from the pilot survey should be viewed as secondary to the comparison of the response behaviours with the two elicitation formats.

5.2.4 Stated preferences & rationality

Stated preference elicitation assumes that respondents are rational, or more specifically, that their preferences are complete and transitive (Lancsar & Louviere 2006). Completeness holds that individuals are able to rank every alternative as more preferred, less preferred or indifferent relative to all other alternatives. This allows for the possibility of utility functions and non-intersecting indifference curves. Transitivity holds that if an individual prefers x to y , and y to z , then they will also prefer x to z . Transitivity rules out the possibility that preferences may 'cycle.' An individual holding some quantity of x , who prefers x to y , and y to z , but intransitively prefers z to x , would in theory be willing to pay some premium to trade x for z , z for y , and y for x . After a cycle of irrational trading, the individual would be back where they started, holding x , but worse-off for having paid a premium at each trade. In the context of revealed preferences – that is, preferences revealed by actual choices – it is assumed that the market will quickly exploit, and thereby correct, irrational preferences. However, as McFadden (1999) notes, there is no market in the context of stated preferences, and therefore no endogenous mechanism to correct irrational preferences. It has therefore been felt necessary to include tests of rationality in stated preference elicitation to prevent irrational preferences from biasing the results. A test of transitivity involves a systematic series of tasks, included among those presented as part of the experimental design, over which respondents are asked to choose between x and y , y and z , and z and x . Respondents whose choices are not consistent with transitivity are flagged as irrational and generally excluded from further analysis.

As recent stated preference research has highlighted, though, seemingly irrational preferences can be based on rational reasons, particularly when respondents may have inferred information that was not included, or intended, as part of the choice task (Miguel et al. 2005; Ryan 2009; Giacomini et al. 2012). This can be exacerbated by the fact that rationality is often judged according to the researcher's expectation of the preferred alternative, which may itself reflect bias or omitted information through poor task design (Lancsar & Louviere 2006). McFadden (1999) argues that to accurately characterise responses as irrational, it is necessary to understand a respondent's perceptions, beliefs,

attitudes, motives and preferences. All of this calls into question the ability of simple tests to distinguish between rational and irrational preferences. In general, Lancsar and Louviere (2006) argue that even though it is clear that not all preferences are rational, “it may not be the case that all preferences labelled as ‘irrational’ are indeed so.” They go on to argue that such irrational responses should not be excluded from interpretation without a very strong theory or empirical evidence to support doing so, and that to do otherwise is to risk imposing the researcher’s *a priori* expectations and preferences on the data. As a practical matter, rationality tests can also add a considerable number of tasks to an experimental design, adding to the time it takes to complete a survey and potentially adversely affecting completion rates and the attentiveness of respondents (Miguel et al. 2005).

Preferences are also generally assumed to be monotonic, stable, and continuous, although these axioms are not essential to rationality (Lancsar & Louviere 2006; Ryan 2009). Monotonicity implies that preferred attributes are ‘goods’ and that more of a good is always preferred to less. Stable, or immutable, preferences imply that if x is preferred to y now, x will continue to be preferred to y in the future, or at least until there is a material change in the relative value of x and y . Finally, continuous, or compensatory, preferences imply that a deterioration in one attribute can be compensated for by an improvement in another. The assumption of compensatory decision making is fundamental to choice-based stated preference elicitation, even though compensatory decision making is cognitively demanding, as it requires decision makers to calculate – implicitly or explicitly – the positive or negative utility derived from the level each attribute, and aggregate utility over each alternative.

Evidence from the psychology literature, though, suggests that decision makers are more likely to be ‘cognitive misers,’ who view decision making a trade-off between the desire to make an optimal decision (the decision benefit) and the desire to minimise the decision cost in terms of cognitive effort or time (Hogarth & Karelaia 2005; Payne et al. 1993; Wright 1975). Compensatory strategies may also require trade-offs that respondents find difficult, or even offensive, particularly if they view the choices to be between rights that should not or cannot be compromised in pursuit of other goals (Bartels & Medin 2007;

Scott 2002). This seems particularly likely in the context of healthcare, where a number of studies have reported that respondents are often reluctant to choose between different patient, viewing such choices as 'playing god' (Cookson & Dolan 1999; Dolan & Cookson 2000; Litva et al. 2002).

Non-compensatory strategies function as 'heuristics' or decision shortcuts that allow decision makers to minimise decision effort and avoid explicit trade-offs between attributes (Gigerenzer & Gaissmaier 2011; Hogarth & Karelaia 2005; Wright 1975). One of the most common heuristics, particularly in the context of a paired stated preference elicitation, is a dominant preference, where a respondent always chooses the alternative with the preferred level of a particular attribute, regardless of the levels of the other attributes (Brandstatter et al. 2006; Gigerenzer & Goldstein 1996; Scott 2002). If the levels of the dominant attribute are equivalent, trade-offs can take place between the other attributes. Lexicographic preferences are a more specific case of dominant preferences where no trade-offs between any attributes takes place. All attributes are ranked by decreasing importance and the decision weight of each attribute is greater than the sum of all weights that come after it. If the levels of the most important attribute are equivalent, then the levels of the second most important attribute are compared, and so on until a preferred alternative is identified (Hogarth & Karelaia 2005; Scott 2002).

Dominant or lexicographic preferences are not irrational as they do not violate the axioms of completeness, transitivity or stability (Mathews et al. 2007; Lancsar & Louviere 2006). Indeed, a dominant or lexicographic preference for aggregate QALY gains is the definition of rationality within the QALY maximising framework. However, such preferences cannot be represented by an indifference curve, and as no trading takes place over some or all attributes, marginal rates of substitution have no meaning (Louviere et al. 2000b; Scott 2002). For this reason, non-compensatory preferences are generally excluded in the interpretation of stated preference data (Lancsar & Louviere 2006; McIntosh & Ryan 2002; Scott 2002). As Lancsar and Louviere (2006) note, though, it may be the case that what appears to be a lexicographic preference may simply be a reluctance to trade over the range of attributes levels in the experimental design. It may also suggest that some or most of the attributes included in the

experimental design are unimportant to some respondents. Therefore, excluding apparently non-compensatory preferences can have the effect of excluding the strongest preferences. The very nature of a fractional factorial experimental design also complicates the interpretation of non-compensatory preferences; as such designs present only a subset of all possible attribute and level combinations, it is not possible to say with certainty that observed instances of non-compensatory decision-making would persist across all possible scenarios (Lancsar & Louviere 2006; Scott 2002).

Despite these limitations in identifying non-compensatory preferences, it was necessary to be able to distinguish respondents with dominant preferences for aggregate QALYs from those willing to sacrifice some QALY gains for equity objectives. Therefore, the identification of QALY maximisers and other non-traders followed an approach outlined by Scott (2002). He acknowledged the difficulty of identifying lexicographic preferences within a fractional factorial design, so to support the characterisation of a respondent as a ‘non-trader’, he applied two criteria. First, dominant preferences were identified as “individuals who always choose the scenario where x_1 is greater than x'_1 , no matter what the level of the other attributes” (Scott 2002). Second, individuals with dominant preferences were classified as non-traders if they also rated that attribute as the most important factor in their decisions in a follow-up rating exercise. This process is described in more detail in the data analysis section.

A test of preference stability was also included, despite the reservations outlined above, in order to compare the DCE and CSPC formats. Miguel et al. (2005) found that increasing choice complexity can lead to an increased incidence of ‘irrational’ responses, which may include unstable or inconsistent preferences. Although it is true that many of individuals flagged as inconsistent may not necessarily be so, a significant difference in the proportions of inconsistent respondents between the DCE and CSPC may be indicative of an overly complex elicitation format.

5.2.5 Data collection

The first phase of the pilot data collection was conducted using informal interviews and focus groups to evaluate the comprehensibility and acceptability of the DCE and CSPC formats. Focus groups were conducted in classes of undergraduate and graduate students in economics and epidemiology, as well as with individual decision makers, healthcare professionals and members of the general public. Focus group participants were presented a short questionnaire with two DCE and two CSPC tasks and asked to offer their feedback on the relative ease or difficulty of understanding the choice task, and their ability to provide a meaningful response. These comments were used to improve the wording and presentation of the tasks. Choice responses collected in this phase were not included in the final dataset. The second phase of the pilot data collection administered full DCE and CSPC questionnaires based on the blocked experimental design detailed above. Responses were elicited from a convenience sample of respondents, including graduate and undergraduate students at Dalhousie University, Halifax, Nova Scotia, Canada, and The University of Sheffield, UK, staff at the Capital District Health Authority in Halifax, Nova Scotia, as well as the general public.

The pilot questionnaires were administered via the internet. Face-to-face administration of stated preference elicitation has significant benefits, including the ability to explain thoroughly the objectives of the survey and to provide timely feedback to respondents (Damschroder et al. 2004). Damschroder et al. (2004) found that respondents to face-to-face surveys were also less likely to provide quick or irrational responses. However, face-to-face administration is costly, time consuming and can often lead to small or selective samples. Relative to less personal elicitation formats, there is also evidence that face-to-face interviews tend to increase 'social desirability' or 'yea saying' biases, where respondents offer the answer they perceive to be socially 'correct' or that will please the interviewer, rather than their true preference (Arrow et al. 1993; Leggett et al. 2003). Although there are also limitations to a web-based approach, the validity of the approach is supported by Damschroder et al. (2004), who found no significant differences in PTO equivalence values elicited using face-to-face and computerised formats. There is also evidence that web-

based administration can minimise social desirability bias in accurately eliciting socially sensitive information (Kreuter et al. 2008). Samples of the pilot DCE and CSPC choice tasks are shown in Appendix 5.1.

Participants were randomised to either a DCE or CSPC questionnaire using a random number algorithm. As each potential respondent followed an online link and was assigned a questionnaire, a record was written to a database indicating the assigned survey. These counts were used as the denominator in calculating the completion rate for each survey. The database counted each time an individual was assigned a questionnaire, but there was nothing to prevent an individual from being counted more than once. For example, individuals who dropped out of an assigned survey but later returned and were re-randomised would have been counted more than once. Additionally, there was no way to ensure that a returning participant was assigned to the same design that they originally started. Therefore, to the extent that some individuals may have been double-counted, completion rates based on these counts were correspondingly underestimated. No demographic information was collected at the time of randomisation, so it was not possible to calculate group-specific completion rates.

Respondents were asked to imagine themselves as a societal decision maker responsible for allocating a fixed budget between two alternative healthcare programs. They were told that both programs had the same cost, and that the budget was large enough to fully fund one program or the other, but not both. The precise budget and the cost of the programs were not specified as realistic program costs are likely to be unfamiliar to respondents and may compromise their ability to make realistic allocations, while trivial sums risk respondents not taking the task seriously (Mullen 1999; Ryan et al. 2001). The concept of cost-effectiveness was not mentioned, but the QALY maximising alternative under an assumption of equal costs will, by extension, also be the more cost-effective alternative, and some respondents may have recognised this fact.

The DCE questionnaire asked respondents to allocate the entire budget to their preferred program, while the CSPC questionnaire asked respondents to allocate budget percentages between the two programs by moving a slider.

Respondents could allocate 100 percent of the budget to program A or program B, or to some combination of the two, including an equal 50-50 split. The number of patients treated in each DCE task was fixed according to the levels in the experimental design, but the number of patients treated in the CSPC tasks was allowed to vary between zero and the maximum level defined by the experimental design in proportion to the budget allocated to each program (e.g. a 25 percent budget allocation meant that 25 percent of the maximum potential number of patients could be treated). It was felt that this would highlight the opportunity cost associated with different budget allocations. The position of the CSPC slider was randomised between each task in order to minimise anchoring and framing effects (Boyle & Ozdemir 2009; Payne et al. 1993).

The CSPC administered here was unique in dynamically linking attribute levels to the budget allocation. Among the CSPC elicitations described in Table 4.2, Schwappach (2003) and Dessler et al. (2010) did not include any attributes that would vary with the relative budget allocation, while Linley and Hughes (2012) skipped the intermediate step of allocating a budget and directly asked respondents how many patients from each of two equally-sized groups they would prefer to treat.¹¹ Linking the number of patients treated – and, indirectly, aggregate QALYs gained – to the relative budget share clearly highlighted the trade-off between the two alternatives. This reality may be obscured in discrete choice tasks as respondents can choose one group without necessarily appreciating that the nature of the task implies that no patients from the other group will be treated.

There is evidence that respondents may choose to avoid difficult choices by selecting an opt-out option, even when one alternative in the choice task may provide greater utility (Ryan & Gerard 2003; Kjær 2005). For this reason, most DCEs in healthcare are based on a forced choice with no opt-out option. However, to minimise dropout from the pilot DCE questionnaire before ratings of difficulty and comprehension could be collected, it was decided to allow

¹¹ This arguably moved the task conceptually closer to a PTO, which asks respondents how many outcomes of type X they would consider equivalent in terms of value to Y outcomes of another kind. However, it is not clear that the final allocation of patients in a CSPC can be interpreted as an indifference point, or in terms of relative value, as it can in a PTO.

respondents to skip tasks without answering. This option was labelled as 'no answer' and these responses were excluded from the analysis. The nature of the CSPC tasks meant that respondents to that questionnaire could indicate equality or indifference between alternatives by selecting an equal 50-50 allocation of the budget. This was taken to indicate sincere indifference, although it is possible that at least some CSPC respondents chose equality in resources as a way to avoid difficult decisions.

Each respondent saw 10 choice tasks, including one repeated task to test preference stability. In this repeated task the position of two alternatives presented in task 3 of each block were reversed and re-presented as task 8. The original choice set was presented as the third task in order to allow respondents to become familiar enough with the tasks to avoid learning effects, and re-presented as the eighth task to allow respondents some time to forget the original choice set, yet not so late as to risk significant fatigue effects. If a respondent's preferences were stable, and if they were paying attention, they should have preferred the same program the same in both choice tasks (Mathews et al. 2007).

Following the choice tasks, respondents were asked to rate the importance of each attribute, including aggregate QALYs and distributional concerns, in their choices on a 0 to 10 scale, and to separately rate the difficulty of understanding and of answering the tasks on 7-point scales ranging from extremely easy to extremely difficult. Respondents were also asked to indicate their gender and age group and to identify themselves as a governmental decision maker or academic expert, a physician, and/or a frequent healthcare user (12 or more healthcare contacts in the past 12 months). These categories were not mutually exclusive, and each respondent could identify as one or more (or none) of these groups.

The questionnaires and the subsequent data analyses were approved by The University of Sheffield Research Ethics Committee, Sheffield UK, and the Capital Health Research Ethics Board, Halifax, Nova Scotia, Canada.

5.2.6 Data analysis

The emphasis in the pilot survey was on identifying a preferred format for the primary elicitation, rather than the estimation of respondent preferences *per se*. Responses from the two surveys were compared on a number of dimensions to assess the difficulty and the acceptability of the two formats, and their ability to elicit valid preference data. These included completion rates, the respondent-rated ease of understanding and answering the questionnaires, preference stability, and the incidence of non-compensatory decision making. A simple analysis of the choice responses was also conducted in order to compare the preference information derived from the two questionnaires. P-values were adjusted for simultaneous comparisons using Hommel's (1988) method,¹² with the exception of p-values on the coefficients in the statistical models of DCE and CSPC choices, which were not adjusted in order to allow for the broadest possible inclusion of potentially explanatory parameters (Hosmer & Lemeshow 2000).

5.2.6.1 Completion rates and respondent-rated difficulty

Differences in questionnaire completion rates and stakeholder and gender proportions were tested using a two-sample Z-test of proportions. Age group proportions were tested using a χ^2 test of independence. On the assumption that the randomisation algorithm assigned an equal proportion of each age, gender and stakeholder subgroup to each questionnaire, differences in these proportions among completed questionnaires were taken to indicate a differential drop-out rate among these groups. The proportions of respondents who indicated that they found the questionnaire 'somewhat difficult' or 'extremely difficult' to understand or to answer were also compared using a two-sample Z-test of proportions.

¹² The more common Bonferroni method for adjusting p-values for n multiple simultaneous comparisons typically sets the acceptable error rate in each comparison (α_n) such that $\alpha_n = \frac{\alpha}{n}$, where α is the overall acceptable error rate (e.g. $\alpha = 0.05$). This approach, though straightforward, is argued to be overly-conservative, often failing to reject the null hypothesis when in fact it is false (Shaffer 1995; Wright 1992). Hommel's method is more complicated but statistically more powerful. Order the hypotheses to be tested by their unadjusted p-value, $p(H_1) \dots p(H_n)$. Let j be the largest integer for which $p(H_{n-j+k}) > \frac{k\alpha}{j}$, for all $k=1, \dots, j$. If no such j exists, reject all the hypotheses; otherwise, reject the hypotheses for which $p \leq \frac{\alpha}{j}$ (Shaffer 1995).

5.2.6.2 Preference stability

Preference stability was measured by including a repeated task in each questionnaire, where the position of the two alternatives presented as task 3 in each design block were reversed and re-presented as task 8. For the purposes of assessing stability, the CSPC budget allocations were transformed to discrete choices on the basis of the alternative to which the majority of resources were allocated. Equal allocations were allowed, but the allocations had to be equal in both tasks in order to be considered consistent. The proportion of consistent responses was compared using a two-sample Z-test of proportions. The statistical significance of the individual differences between the initial and repeated budget allocations in the CSPC questionnaires was tested using a paired t-test.

5.2.6.3 Dominant preferences

As Scott (2002) noted, lexicographic preferences are rarely identifiable in the context of a stated preference elicitation, but it is generally possible to identify dominant preferences. To test for such preferences, a set of flags was created for each alternative in each choice task. These flags indicated whether or not an alternative presented the most preferred, or dominant, level of each attribute. For example, based on evidence of public support and an ethical justification for prioritising more severely ill patients from the empirical ethics review, if one alternative presented patients in a more severe initial health state, that alternative was flagged as 'best' (from the perspective of the respondent) in the initial utility attribute; the corresponding attribute flag for the paired alternative was set to zero. Similarly, if one alternative was associated with greater life year gains than the other, that alternative was flagged as best in that attribute. There were a total of seven flags for each alternative: age, initial utility, initial life expectancy, final utility, life years gained, (potential) number of patients treated and (potential) number of QALYs. CSPC responses were transformed to discrete choices on the basis of the program to which the respondent allocated the majority of the budget, and the flags were set based on the potential number of patients that could be treated and the potential number of QALYs gained if 100% of the budget were allocated to that alternative. CSPC

alternatives that received a 50% budget allocation were flagged as ‘not chosen’ (i.e. both alternatives were assigned a choice flag of zero) as neither alternative was prioritised, but the impact of counting such allocations as prioritising *both* attributes was also tested.

The absolute value of the correlation between choice and each attribute flag, measured by Kendall’s tau (Herve 2007), was taken to represent the degree of that attribute’s dominance in each respondent’s choices. A respondent that always chose the alternative with, for example, the younger patients, would have a choice correlation coefficient of 1.0 with the age attribute. Which end of each attribute scale the respondent considered ‘best’ was not critical, as in this example correlation would -1.0 if they always chose the alternative with the older patients. It is important to note, though, that this approach to identifying dominant preferences only holds where preferences are monotonically increasing or decreasing over the attribute, as was assumed here.

As respondents saw only a subset of possible scenarios, it was not possible to say that a perfect correlation between a respondent’s choices and the level of a particular attribute would necessarily hold across all possible scenarios (Scott 2002). Therefore, to support the identification of non-traders, each respondent’s self-rated attribute importance scores were converted to rankings, and individuals with a perfect choice-attribute correlation who also rated that attribute as most important were considered to have a dominant preference for that attribute. The proportion of non-traders was compared across the two questionnaires using a two-sample Z-test. A very high incidence of non-compensatory preferences in a particular questionnaire may invalidate the interpretation of the responses, while a significant difference between questionnaires may reflect excessive task complexity and a corresponding degree of simplification in one of the formats. Similarly, CSPC respondents who allocated every budget so as to equalise resources, the number of patients treated, or the aggregate QALYs gained in each group were characterised as strict (non-trading) egalitarians if they also ranked the distribution of resources as the most important factor in their choices.

With respect to preferences for aggregate QALYs, recognise that in the context of equal program costs, a dominant preference for greater aggregate

QALYs is also a preference for the more cost-effective alternative. However, in holding costs constant it was not possible to distinguish between a preference for the more cost-effective alternative, and a dominant preference for aggregate QALYs that may have held regardless of relative cost. A dominant preference for aggregate QALYs was therefore necessary but not sufficient to confirm support for the principles of QALY maximisation. To test whether one format was associated with a greater preference for aggregate QALYs, even if these preferences were not necessarily dominant, the mean number of QALY maximising choices made by respondents to the two questionnaires was compared using a two-sample t-test.

5.2.6.4 Choice analysis

Given the limited degrees of freedom available in each block of the experimental design, the choice models assumed monotonic preferences and only estimated linear main effects. The QALYs gained attribute, as a linear combination of initial and final utility, life years gained and number of patients treated, was excluded from the analyses in order to avoid collinearity. Responses to the repeated task were excluded to avoid double counting, as were ‘no answer’ responses from the DCE questionnaire. All respondents were included in the analysis, including those identified as non-traders. The analyses were performed with R, version 2.15.2 (R Core Team 2013) using the mlogit (Croissant 2012), censReg (Henningsen 2012) and plm (Croissant & Millo 2008) packages.

CSPC responses were analysed using a double-bounded tobit model to account for the censored dependent variable. Although the previous chapter mentions the potential advantages of a latent class model, a random effects specification was adopted for simplicity and parsimony in the pilot analysis.

$$\begin{aligned} \Delta BUD_{it}^{B-A} = & \alpha + \beta_1 \Delta Age_{it}^{B-A} + \beta_2 \Delta UO_{it}^{B-A} + \beta_3 \Delta LE_{it}^{B-A} \\ & + \beta_4 \Delta U1_{it}^{B-A} + \beta_5 \Delta LYg_{it}^{B-A} + \beta_6 \Delta Pats_{it}^{B-A} + \mu_i \quad (5.8) \\ & + \varepsilon \end{aligned}$$

The response variable (ΔBUD_{it}^{B-A}) was the budget allocated to Program B less the budget allocated to Program A by respondent i in task t . If 100% of the budget was allocated to Program B, $\Delta BUD_{it}^{B-A} = +100$; if 100% was allocated to

Program A, $\Delta BUD_{it}^{B-A} = -100$; if the budget was allocated 50%-50%, $\Delta BUD_{it}^{B-A} = 0$. Similarly, the parameters were the differences in the continuous attribute levels between Program B and Program A. The β 's represented the change in latent utility associated with a one-unit increase in the level of an attribute, μ_i was an individual-specific error term, and ε was a stochastic error term. To be as consistent as possible with the CSPC analysis, DCE responses were modelled using a binary random effects probit, where the parameters were defined as in equation 5.8, but the response variable was a 0, 1 flag indicating whether or not alternative B was chosen. As noted in section 4.5, there are limitations to a random effects specification, but it was felt to be sufficient for the pilot elicitation given its emphasis on the response characteristics of the two questionnaire formats rather than respondent preferences *per se*, and its limited degrees of freedom. The DCE and CSPC models were compared in terms of the relative contribution of each attribute to overall utility, or the relative importance of each attribute (Orme 2006a), as well as the marginal rates of substitution between individual life years gained and the other parameters in each model.

In the DCE, each attribute's contribution to systematic utility was calculated based on the most preferred and least preferred level of attribute x :

$$\Delta v(x) = (\beta x)^{max} - (\beta x)^{min} \quad (5.9)$$

Where $(\beta x_i)^{max}$ was the utility associated with the *most* preferred level of attribute x , $(\beta x_i)^{min}$ was the utility associated with the *least* preferred level of attribute x , and $\Delta v(x)$ was the net difference in utility. This attribute-specific contribution was then divided by the difference in overall utility between the 'best' scenario (v^{max}), based on the most preferred levels of all the statistically significant attributes in the model, and the 'worst' scenario (v^{min}), based on the least preferred levels of all the attributes:

$$\text{Relative Importance of } x = \frac{\Delta v(x)}{v^{max} - v^{min}} \quad (5.10)$$

Where $\Delta v(x)$ is defined as in equation 5.9 above. The calculation was essentially the same for the CSPC, except that the x 's represented the smallest and largest *differences*, rather than absolute levels, and $\Delta v(x)$ was the overall difference in latent utility.

As the levels of each attribute are measured on scales with different origins and different units, this means that the attribute coefficients within each model cannot be compared unless they are transformed to some common scale (Lancsar et al. 2007). The coefficients were transformed on the basis of marginal rates of substitution (MRS), using individual life years gained as the numeraire:

$$\text{MRS} = \frac{\beta_x}{\beta_{LYg}} \quad (5.11)$$

Where β_x is the coefficient on attribute x , and β_{LYg} is the coefficient on the individual life years gained attribute. MRS represents the number of individual life year gains that respondents would, in theory, be willing to sacrifice in return for a 1-unit change in the level of attribute x . A statistically significant and negative MRS indicated a preference for a lower level of an attribute, while a statistically significant and positive MRS indicated a preference for a higher level.

5.3 Results

Data collection for the pilot survey ran from March to May 2011, and a total of 604 individuals began a questionnaire: 348 (58%) were randomised to the CSPC questionnaire and 256 (42%) were randomised to the DCE questionnaire. Participants were initially allocated between the two questionnaires on an equal basis, but to compensate for lower observed completion rates among participants allocated to the CSPC in the early stages of data collection this was adjusted to allocate an arbitrary 60 percent of participants to the CSPC questionnaire. Completion rates and respondent characteristics are shown in Table 5.2.

Table 5.2: Respondent characteristics by questionnaire

| | DCE (%) | CSPC (%) | p-value | Adjusted-p | Sig |
|--|---------------|---------------|---------|------------|-----|
| Overall completion rate | 154/256 (60%) | 150/348 (43%) | <0.001 | <0.001 | |
| Self-identified stakeholders, N (%) | | | | | |
| Decision maker | 33 (21%) | 18 (12%) | 0.04 | 0.20 | |
| Doctor | 35 (23%) | 35 (23%) | 1.00 | 1.00 | |
| Frequent user | 14 (9%) | 18 (12%) | 0.52 | 1.00 | |
| Demographics, N (%) | | | | | |
| Female | 113 (74%) | 107 (71%) | 0.77 | 1.00 | |

| | | | | |
|--|------|------|------|------|
| Mean age* | 31.5 | 33.2 | 0.65 | 1.00 |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+' | | | | |

* Mean age was calculated using the mid-point of the different age groups; the p-value was based on a χ^2 test of independence

A significantly greater proportion of individuals completed the DCE questionnaire compared with the CSPC questionnaire, suggesting that the CSPC may have been less acceptable to respondents in some respects. There were no significant differences in the gender distribution or mean age, or in the proportion of respondents who identified themselves as doctors or frequent healthcare users. However, a lower proportion of CSPC respondents identified themselves as a government decision maker or academic expert, suggesting a higher drop-out rate among this group in the CSPC relative to the DCE, although this difference was not statistically significant after adjusting for multiple comparisons. Conversely, a slightly greater proportion of frequent healthcare users completed the CSPC questionnaire, although again this difference was not statistically significant. As frequent users may be more likely to be chronically ill than the other respondent subgroups – and therefore also more likely to feel that they may be viewed as a less preferred group by the larger society – it is possible that these respondents may have preferred the CSPC, with its ability to reserve some resources for less preferred groups, to a greater degree than the other subgroups. Note that these groups were not mutually exclusive as respondents could identify as belonging to more than one stakeholder group.

The CSPC completion rate was similar to that reported by Ratcliffe (38%) in her application of CSPC (Ratcliffe 2000), but the DCE completion rate was lower than the 77 percent completion rate reported by Norman et al (2013) and the 75 percent reported by Shah et al. (2012).

5.3.1 Respondent-rated difficulty

As shown in Table 5.3, there was no significant difference between the two surveys in the proportion that rated the tasks ‘somewhat difficult’ or ‘extremely difficult’ to understand among all respondents who submitted a questionnaire.

Table 5.3: Respondents rating the questionnaires ‘somewhat difficult’ or ‘extremely difficult’ to understand

| | DCE (%) | CSPC (%) | p-value | Adjusted-p | Sig |
|---|----------------|----------------|---------|------------|-----|
| All respondents | 19/154 (12.3%) | 19/150 (12.6%) | 1.00 | 1.00 | |
| Decision maker | 5/33 (15.2%) | 5/18 (27.8%) | 0.47 | 1.00 | |
| Doctor | 6/35 (17.1%) | 5/35 (17.1%) | 1.00 | 1.00 | |
| Frequent user | 1/14 (7.1%) | 2/18 (11.1%) | 1.00 | 1.00 | |
| Significance codes: <0.001= ‘***’ <0.01= ‘**’ <0.05= ‘*’ <0.10= ‘+’ | | | | | |

Among stakeholder subgroups, a greater proportion of decision makers found the CSPC tasks difficult to understand compared to the DCE, although this difference was not statistically significant, even before adjusting for multiple comparisons. Likewise, Table 5.4 suggests that there were no statistically significant differences overall or by respondent subgroup in the perceived difficulty of answering the tasks, although a greater proportion of all decision makers (across both questionnaires) reported the tasks to be difficult to answer (76.5%) relative to all other respondents excluding decision makers (63.2%). This difference, though, was not statistically significant (p=0.10, adjusted-p=0.49).

Table 5.4: Respondents rating the questionnaires ‘somewhat difficult’ or ‘extremely difficult’ to answer

| | DCE (%) | CSPC (%) | p-value | Adjusted-p | Sig |
|---|-----------------|----------------|---------|------------|-----|
| All respondents | 100/154 (64.9%) | 99/150 (66.0%) | 1.00 | 1.00 | |
| Decision maker | 25/33 (75.8%) | 14/18 (77.8%) | 1.00 | 1.00 | |
| Doctor | 20/35 (57.1%) | 21/35 (60.0%) | 1.00 | 1.00 | |
| Frequent user | 7/14 (50.0%) | 11/18 (61.1%) | 1.00 | 1.00 | |
| Significance codes: <0.001= ‘***’ <0.01= ‘**’ <0.05= ‘*’ <0.10= ‘+’ | | | | | |

As these difficulty ratings were limited to those respondents who submitted a completed questionnaire, they may be biased downwards as individuals who found the surveys exceedingly difficult are more likely to have dropped-out before completion. However, the proportions reporting the questionnaires to be somewhat or extremely difficult to understand or to answer were similar to the proportions reported by other authors using DCE or CSPC methods to elicit societal preferences. Green and Gerard (2009) reported that

40% of respondents found the DCE difficult to understand and 68% found it difficult to answer, while Ratcliffe (2000) reported that 41 percent of respondents found her CSPC moderately or very difficult to complete, and Schwappach (2003) reported that 52 percent of respondents found his CSPC quite or very difficult.

5.3.2 Preference stability

In the DCE survey, 148 out of 154 respondents (96%) preferred the same program (including 3 consistent 'no answers') in the original and the repeated task. After transforming budget allocations to discrete choices on the basis of the program to which the majority of the budget was allocated, 119 out of 150 CSPC respondents (79%) allocated the majority of the budget to the same program or preferred an equal allocation of resources in both tasks. The difference in the proportion of respondents that chose the same alternative in the repeated task was significantly greater in the DCE questionnaire than the CSPC ($p < 0.001$). While this suggests that not all respondents had stable preferences, or that not all respondents were not paying attention to their choices, some of the observed inconsistencies may be explained by respondents adopting an egalitarian perspective on their choices: if a respondent remembered prioritising a group with the same characteristics in the original task, they may have wanted to 'even out' the allocation of resources by prioritising the other group in the repeated task.

When the individual differences between the specific budget allocations in the original and repeated CSPC task were considered, the mean budget allocation to program B in the original task was 27 percent compared to 19 percent in the repeated task, for a net difference of -8 percent ($p < 0.001$, adjusted- $p < 0.001$). It is worth noting, however, that the mode budget difference – accounting for 18 percent of the paired responses from the original and repeated task – was zero, indicating the same allocation in both tasks.

5.3.3 Dominant preferences and non-trading behaviour

Excluding three individuals who always chose 'no answer' in the DCE, 9 percent of DCE respondents (14/151) and 5 percent of CSPC respondents (7/150) always chose the alternative with the preferred level of a particular attribute. This difference was not statistically significant ($p=0.27$). The attribute most frequently perfectly correlated with choice in the DCE was final health state (9/14), and in the CSPC it was individual life years gained (6/7). Among the respondents with at least one perfectly correlated attribute, 7 DCE (5%) and 3 CSPC (2%) respondents also ranked that attribute as the most important factor in their choices, which was taken as confirmation of a dominant preference. Again, this difference was not statistically significant ($p=0.34$). When equal CSPC budget allocations were counted as prioritising the dominant attribute, the proportion of respondents with perfect choice-attribute correlations increased to from 5 to 11 percent (16/150), and the proportion with a confirmed dominant preferences increased from 3 to 5 percent (8/150), but the differences between the CSPC and DCE were still not significant ($p=0.83$ and 0.99 , respectively).

Three additional DCE respondents had a perfect correlation between choice and total patients treated, but due to a coding error in the database attribute importance ratings were not recorded for the number of patients treated attribute, and it was not possible to confirm a dominant preference for these respondents. If all 3 had ranked this attribute as the most important factor in their choices it is possible that up to 12 DCE respondents (8%) may have had a dominant preference, although this would not change the statistical insignificance of difference between the DCE and CSPC ($p=0.49$).

With specific reference to aggregate QALY gains, only one respondent, from the DCE questionnaire, chose the QALY maximising alternative in every task. This individual also ranked QALYs as the most important attribute, confirming a dominant preference for aggregate QALYs. On average, DCE respondents chose the QALY maximising alternative in 6.3 out of 10 tasks, compared to 5.4 tasks out of 10 among CSPC respondents ($p < 0.001$). Both rates were slightly but significantly greater than the 5 choices out of 10 that would be expected by chance alone, given that one alternative in each choice pair maximised QALYs gained (adjusted- $p < 0.001$ in both comparisons).

However, this relatively low rate of prioritising the QALY-maximising alternative in either formats appeared to offer little support for QALY maximisation as a societal decision rule.

The proportion of respondents with a confirmed dominant preference in both of the questionnaires was less than the 45 percent reported by Scott (2002) and the 19 percent by Norman et al. (2013) using DCE methods, and the 5.7 percent reported by Chan (2006) using a CSPC. However, it was greater than the 2 percent reported by Schwappach (2003) and the 0.3 percent reported by Ratcliffe (2000), both of whom also used a CSPC.

5.3.4 Choice analysis

Although respondents to the DCE and CSPC did not see the same choice sets, there were only weak correlations between the block indicator and the specific attribute levels, ranging from -0.14 to 0.16, suggesting that there was no systematic bias in the attribute levels presented in the two questionnaires.

Appendix 5.2 presents the model coefficients and p-values from the DCE and CSPC models, along with marginal rates of substitution (MRS) and relative attribute importance weights and rankings. Initial life expectancy and the number of patients treated were not significant at a 0.10 threshold in the initial DCE probit model, and it was re-estimated excluding these attributes. All six attributes were significant in the CSPC model. The direction of preferences was consistent between the two models: negative coefficients on age and initial utility suggested that younger and more severe patients were preferred, while positive coefficients on individual life years gained and final utility suggests respondents preferred greater individual life year gains and better final health states. CSPC respondents also preferred larger patients groups and patients with greater initial life expectancy, while these attributes were not significant in the DCE.

As illustrated in Figure 5.1 below, within the range of the attributes tested, final health state was the single most important attribute in both models, with relative importance weights of close to 50 percent in both questionnaires, but the rankings diverged for the other attributes. The next most important attribute in the CSPC was initial health state, where respondents had a preference for patients in more severe initial health states, while DCE respondents gave more importance to individual life year gains. CSPC respondents also had a preference for larger patient groups, while the number of patients was not statistically significant in the DCE. This result is notable as it may be reflective of a ‘prominence effect,’ by which respondents may become more sensitive to a quantity when it is harder for them to ignore (Baron & Greene 1996; Fischer et al. 1999). In this case, the number of patients treated changed as CSPC respondents moved the budget slider, potentially highlighting this attribute and leading CSPC respondents to give it more weight in their choices.

5.3.5 CSPC budget allocations

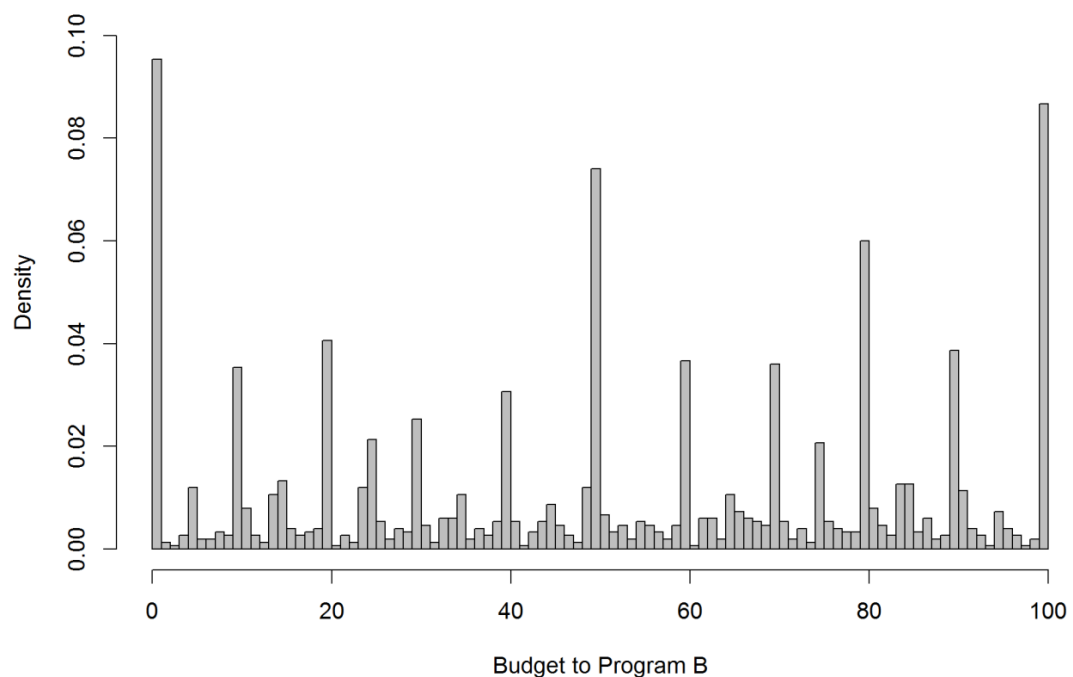
Figure 5.2 shows that the modal CSPC allocation (18% of all responses) maximised the budget allocation to program A (0 percent to program B) or program B (100 percent to program B), while 7 percent of responses equalised

Figure 5.1: Attribute relative importance by format



the budget between the two alternatives. At the respondent level, 2 percent of respondents (3/150) maximised the budget in every task, and 11 percent (16/150) maximised the budget in 5 or more of their 10 choices. No respondents equalised budgets, patients or QALYs in more than 5 of their choices, and there were no significant differences in the proportion of decision makers who equalised or maximised relative to all other respondents in the survey.

Figure 5.2: Pilot CSPC budget allocations



The absence of any respondents who equalised allocations in every task was in contrast to the results of CSPC elicitation reported by Ratcliffe (2000) and Chan et al. (2006), who both found that 2 percent of respondents equalised resources in every task, and to Schwappach (2003), who found that 11 percent of respondents equalised resources in every task. Schwappach also reported that only 3 percent of all allocations maximised the budget allocation to one group or the other, sharply contrasting with the much higher proportion observed here. In general, the low rates of equalising behaviour suggest that respondents were not using equal budget allocations as a way to avoid making difficult allocation choices.

5.4 Identifying a preferred elicitation format

In order to identify a preferred format for the primary preference elicitation, DCE and CSPC questionnaires were compared in terms of completion rates, difficulty ratings, preference stability, the incidence of non-trading behaviours (including strict QALY maximisation), and attribute importance weights. The results of these comparisons, though, did not seem to identify a clearly superior alternative. The clearest advantage was in terms of completion rates, where a significantly greater proportion of assigned DCE participants completed a questionnaire compared to assigned CSPC participants. This suggested that participants found the DCE more acceptable than the CSPC in some respects, even though the difficulty ratings of the two questionnaires, in terms of understanding as well as answering, were virtually identical. This appeared to undermine Schwappach and Strasmann's (2006) suggestion that CSPC may be more acceptable to respondents in a healthcare context given its ability to avoid extreme distributions, although there may have been an element of this in the higher completion rate observed with the CSPC relative to the DCE among frequent healthcare users. It also offered little support for Swallow et al.'s (2001) suggestion that dichotomous choice tasks may leave respondents dissatisfied with the limited information they are allowed to provide. It was somewhat surprising to note that decision makers, who might be expected to be more familiar with abstract choice tasks, expressed the greatest difficulty in understanding the CSPC, and as a group, reported the greatest difficulty in answering the tasks in both questionnaires. This group was also less likely to complete the CSPC questionnaire compared to the DCE questionnaire.

Respondents to the DCE questionnaire were significantly more consistent in preferring the same alternative in the repeated task, suggesting greater preference stability – or at least greater attentiveness – among these respondents. However, the choice sets that were arbitrarily chosen for the repeated task may have contributed to the observed stability. Ninety-five percent of respondents preferred the same alternative in the original DCE task, compared to only 77 percent in the original CSPC task, and the near unanimity of choice in the DCE task suggests that one alternative was an 'obvious' choice and therefore an overly easy test of preference stability. An ideal test would have presented two

alternatives with roughly equal choice probabilities, although it was not possible to predict these probabilities prior to the elicitation.

Respondents to the CSPC appeared slightly – although not significantly – less likely to demonstrate non-trading behaviours in the form of a non-compensatory dominant preference for a particular attribute than respondents to the DCE. This is consistent with the notion that the more competitive nature of the ‘pick one’ DCE task may tend focus attention on a single attribute to a greater degree than a relatively more reflective task such as the CSPC (Huber 2009). It also suggests that respondents to the more cognitively demanding CSPC were no more likely to resort to a simplifying heuristic such as a dominant or lexicographic preference than respondents to the DCE.

With specific reference to QALY maximising behaviour, only one respondent to either questionnaire had a dominant preference for aggregate QALYs gained, and there was a relatively low overall proportion of QALY maximising choices in either questionnaire. This is particularly noteworthy in the CSPC, where aggregate QALYs gained changed along with total patients treated as respondents moved the budget slider. The statistical significance of the number of patients treated attribute in the CSPC budget allocations, in contrast to its insignificance in the DCE choices, may suggest the possibility of a prominence effect associated with this dynamic link in the CSPC tasks. In light of qualitative evidence that suggests respondents to stated preference elicitation often reduce abstract, macro-level allocation problems to more comprehensible two-person analogies (Giacomini et al. 2012; Ryan 2009), an effect that ‘nudges’ respondents to account for the macro- or societal-level implications of their choices might be an advantage in maintaining the intended perspective of a societal preference elicitation (Fischer et al. 1999; McQuillin & Sugden 2012). However, if there was indeed such an effect associated with the number of patients treated, it did not appear to carry over to aggregate QALYs gained. Furthermore, the evidence for a prominence effect in the CSPC must be interpreted cautiously, given the relatively small sample sizes and the fact that the respondents to the two questionnaires did not see the same experimental design. It is not possible to say with certainty, therefore, that the observed differences in attribute importance weights were due to the elicitation formats

themselves, and not simply due to differences in the choice sets presented to respondents.

Notwithstanding the low proportion of strict QALY maximisers in the CSPC, there was an unexpected willingness among CSPC respondents to maximise budget allocations that challenged previous studies that found a general aversion to extreme distributions (Ratcliffe 2000; Schwappach 2003). Although equal budget allocations were relatively common among the CSPC responses, the weak overall preference for an egalitarian distribution of resources seems noteworthy, as in the absence of any obvious rationale for a particular budget allocation respondents could have been expected to use an equalising allocation (of resources, patients or outcomes) as a heuristic for a 'fair' allocation. Indeed, as Culyer (2001b) notes, equity and fairness are generally held to imply equality in *something*. Instead, consistent with a random utility theory interpretation of the results, it appeared that respondents chose allocations were based on a view of the relative utility of the paired alternatives. It also suggested that CSPC respondents to this survey did not tend to use equal budget allocations as a way to avoid difficult decisions.

Overall, the DCE questionnaire appeared to be better at eliciting responses, as more participants completed it compared to the CSPC questionnaire. However, as judged by the statistical significance of the different attributes in the choice models, respondents to the CSPC appeared to incorporate more of the attributes presented in each alternative into their choices compared to DCE respondents. Consistent with a narrower focus, DCE respondents also appeared more likely to have a dominant preference for a single attribute in their choices. This suggests a possible tension between the quantity and the quality of responses elicited by the two stated preference methods, and that ultimately, potentially richer preference data with CSPC must be weighed against better completion rates and preference consistency ('respondent efficiency' (Severin 2001)) with DCE. As neither format distinguished itself as clearly superior, it was decided to proceed with both formats for the primary elicitation. This allowed for further exploration of their response behaviours and a fuller comparison of the preference weights using more sophisticated models

based on an optimal experimental design. The methods and results of these elicitations are described over the next four chapters.

Appendix 5.1: Sample DCE and CSPC choice tasks

Sample discrete choice experiment

Both Program A and Program B have the same total cost. A fixed budget has been set aside to fund one of these programs, but it is not large enough to fund both of them. The budget will only fund healthcare and cannot be used for research that may improve a patient's condition in the future. After considering the characteristics of each program, please indicate which program you would prefer to fund. Click on the [blue](#) text to get a more complete explanation of each attribute.

| Program A | Attributes | Program B |
|-------------------------|---|------------------------------|
| 70 years old | Average age of patients | 10 years old |
| 1 out of 10 | Quality-of-life without/before treatment | 9 out of 10 |
| 10 years | Life expectancy without/before treatment | 1 month |
| 1 out of 10 [No change] | Quality-of-life with treatment | 5 out of 10 [4 levels lower] |
| 10 additional years | Change in life expectancy with treatment | 1 additional year |
| 5,000 | Number of patients that could benefit | 500 |
| 5,000 | Total quality-adjusted life years gained with treatment | 250 |

No answer
 I would prefer to fund Program A
 I would prefer to fund Program B

Sample constant sum paired comparison task

Both Program A and Program B have the same total cost. A fixed budget has been set aside to fund these programs, but it is not large enough to entirely fund both. The money in the budget can only be used to fund these programs and it cannot be used for research that may improve patients' condition in the future. Using the slider below you can split the budget any way you like between both programs, including 100% to Program A or Program B. As you adjust the slider, the total number of patients that can be treated and the total number of quality-adjusted life years generated by each program will change along with the budget. Click on the [blue](#) text to get a more complete explanation of each attribute.

| Program A | Attributes | Program B |
|-------------------------|---|-------------------------|
| 10 years old | Average age of patients | 40 years old |
| 1 out of 10 | Quality-of-life without/before treatment | 9 out of 10 |
| 5 years | Life expectancy without/before treatment | 1 month |
| 1 out of 10 [No change] | Quality-of-life with treatment | 9 out of 10 [No change] |
| 1 additional year | Change in life expectancy with treatment | 10 additional years |
| 335 | Number of patients that could be treated | 825 |
| 33 | Total quality-adjusted life years gained with treatment | 7,425 |

Percent of budget to Program A

67%

Use this slider to shift the total budget between Programs A and B



Percent of budget to Program B

33%

Respondent attribute rating task

Please indicate how important each of the factors listed below were to you in deciding how to allocate public healthcare resources. 0 stars means a factor had no bearing on your decision and 10 stars means it was the most important attribute in your decision. Click on the [blue](#) text for a more complete definition of the attribute.

Clicking on a star will fill in all the stars to the left

- The average age of the patients ☆☆☆☆☆☆☆☆☆☆
- [Quality-of-life without/before treatment](#) ☆☆☆☆☆☆☆☆☆☆
- [Life expectancy without/before treatment](#) ☆☆☆☆☆☆☆☆☆☆
- Change in quality-of-life with treatment ☆☆☆☆☆☆☆☆☆☆
- [Change in life expectancy with treatment](#) ☆☆☆☆☆☆☆☆☆☆
- The number of patients in each group that could be treated ☆☆☆☆☆☆☆☆☆☆
- [Total quality-adjusted life years gained with treatment](#) ☆☆☆☆☆☆☆☆☆☆
- [Making sure each group has an equal share of healthcare resources](#) ☆☆☆☆☆☆☆☆☆☆

Attribute descriptions

- **Quality-of-life** refers to how well/sick a patient is before treatment and is measured on an imaginary 0 to 10 scale, where 0 means death and 10 is perfect health. At level 1, patients have severe problems with pain and mobility and they are unable to perform their usual activities. At level 5, patients have moderate problems with pain and mobility and they can only participate in some of their usual activities. At level 9, patients have very minor or no problems with pain and mobility and they are able to participate in all their usual activities.
- **Life expectancy** refers to how long the average patient will live from today.
- **Treatment** is a hypothetical drug or procedure that could improve a patient's quality of health, life expectancy or both.
- **Change in health with treatment** refers to the improvement in quality-of-life a patient gets from treatment. It is measured on the same scale mentioned above, where 0 means death and 10 is perfect health. Because cancer treatment can involve harsh chemotherapy drugs and radiation, it is possible that treatment could reduce a patient's quality-of-life in order to extend the length of their life.
- **Change in life expectancy** refers to how many additional years of life a patient will gain from treatment. A patient's total life expectancy with treatment would be their initial life expectancy without treatment plus their change in life expectancy.
- **Total life years gained** measures the total number of additional number of years lived, adding across all patients in the program. For example, if 5 patients live an additional 5 years each, the total life years gained is 25.

- **Quality-adjusted life years (QALYs)** are a measure of the total benefit of a health program. It considers changes in the length of life, change in quality of life and the number of patients that can be treated. If a person spends 1 additional year at health level 10 out of 10 (perfect health), they would count for 1 QALY. If they spent 1 additional year at health level 5 out of 10 (i.e. half as good as level 10, or perfect health) they would count for 0.5 QALYs. Total QALY gains are calculated as the change in life expectancy \times the change in quality \times the number of patients treated.
- The **share of healthcare resources** refers to the portion of the healthcare budget that one patient group receives compared to the other group. Ensuring each group gets an equal share means you believe that each group should always receive an equal share of the budget regardless of their age, quality-of-life, life year gains, QALYs, or any other of the characteristics listed above.

Appendix 5.2: DCE & CSPC choice model coefficients, marginal rates of substitution and importance ranks

| DCE | | | | | | | |
|---------------------------|----------|------------|--------------|---------|-----|----------|----------------------|
| Attribute | Estimate | Std. Error | Coeff of Var | p-value | Sig | MRS(LYg) | Rel. Importance Rank |
| Constant | 0.158 | 0.048 | 0.303 | <0.001 | *** | | |
| Δ Life years gained | 0.086 | 0.011 | 0.133 | <0.001 | *** | 1.00 | 0.321 2 |
| Δ Age | -0.008 | 0.001 | 0.147 | <0.001 | *** | -0.09 | 0.202 3 |
| Δ Initial utility | -0.058 | 0.009 | 0.154 | <0.001 | *** | -0.67 | 0.019 4 |
| Δ Initial life expectancy | | | | | | | |
| Δ Final utility | 0.138 | 0.012 | 0.086 | <0.001 | *** | 1.61 | 0.458 1 |
| Δ Patients treated / 1000 | | | | | | | |
| CSPC | | | | | | | |
| Attribute | Estimate | Std. Error | Coeff of Var | p-value | Sig | MRS(LYg) | Rel. Importance Rank |
| Constant | -32.255 | 3.891 | 0.121 | <0.001 | *** | | |
| Δ Life years gained | 7.298 | 0.420 | 0.057 | <0.001 | *** | 1.00 | 0.101 4 |
| Δ Age | -0.137 | 0.053 | 0.390 | 0.010 | * | -0.02 | 0.013 6 |
| Δ Initial utility | -18.480 | 3.750 | 0.203 | <0.001 | *** | -2.53 | 0.228 2 |
| Δ Initial life expectancy | 2.995 | 0.411 | 0.137 | <0.001 | *** | 0.41 | 0.046 5 |
| Δ Final utility | 40.101 | 3.583 | 0.089 | <0.001 | *** | 5.49 | 0.496 1 |
| Δ Patients treated / 1000 | 16.705 | 1.544 | 0.092 | <0.001 | *** | 2.29 | 0.116 3 |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

MRS(LYg)=marginal rate of substitution using life years gained as the numeraire

Chapter 6: Primary data collection methods and sample characteristics

The objective of the primary survey was to estimate preference weights for the attributes identified in the empirical ethics review of Chapter 3 from a representative sample of the Canadian public, and the intention was to use the elicitation method that performed best in the pilot elicitation. However, the pilot survey did not identify a clearly preferred method, and in fact raised a number of interesting methodological issues that would benefit from further exploration. These included the possibility of a prominence effect around the number of patients treated, and questions around the relative stability of preferences with the two methods. Therefore, it was decided to proceed with both methods for the primary survey of preferences. Secondary objectives of this survey included a more detailed comparison of the properties of the two elicitation methods, taking advantage of the larger and more representative sample, and to compare the preferences of the general public with those of decision-making agents responsible for making prioritisation decisions on behalf of the public.

As described by Coast (2001a), agents in this context are part of an implicit principal-agent relationship with the public, where the public may feel that they are ill-informed about their preferred allocation of healthcare resources and relies on agents to make allocation decisions on their behalf. Culyer (1989), in the context of the traditional doctor-patient relationship, saw the ideal agency relationship as one where the agent makes the decision that the 'client' would have if they had the same knowledge and information. As argued in Chapter 2, however, the goodness of any particular allocation of societal resources is a subjective truth, and as such, no special knowledge or unique objectivity is

required or even necessarily desirable, as such restrictions would be counter to Sen's (2011) view of the importance of the 'universality of inclusion' in societal decision-making. By this view, agent preferences should reflect those of the public. Brouwer et al. (2008), though, suggest that the responsibility of a societal decision maker is not to reflect how citizens *would* act, but rather how they *ought* to act. By this view, agent preferences should be expected to diverge from those of the general public, perhaps in terms of a greater emphasis on QALY maximisation.

This chapter describes the methods for the primary preference elicitation surveys, which drew heavily on the methods used in the pilot survey, and summarises the characteristics of the survey respondents. The development of the experimental design, the target sample populations, and the format of the DCE and CSPC choice and rating tasks are described in section 6.1. Section 6.2 describes the master experimental design administered to respondents, and reports on the correlations between each of the attributes in the design. Some correlation is inevitable given the fractional factorial design, but ideally these correlations should be relatively small. A descriptive summary of the survey, including the total number of respondents, their characteristics, and the total number of choices made by respondents is presented in section 6.3, along with their attitudes towards healthcare rationing. These attitudinal questions included their agreement or disagreement with the need for rationing, their support for different stakeholder groups in rationing decisions, and their comfort with having their preferences used in priority setting decisions. The implications of these attitudes for a democratic or Communitarian approach to priority setting are discussed in section 6.4.

6.1 Survey methods

The survey methods were based closely on the on the pilot methods, although there were some differences in the experimental design and data collection. There were also some additional tests of validity and rationality included in each questionnaire, along with some attitudinal rating tasks. The differences between the two methodologies are detailed below.

6.1.1 Experimental design

The experimental design used the same attributes as the pilot survey. Each alternative was described in terms of the average age of patients, their health state (utility) and life expectancy without treatment, health state and life years gained with/after treatment, the number of patients that could be treated in each group, and aggregate QALYs gained, which was calculated as a product of the other attribute levels. The only change from the pilot survey was in the number of patients treated attribute: the levels were changed from 500, 2000 and 5000, to 100, 2500 and 5000 to have a slightly greater range and to be more symmetrical around the middle level.

Table 6.1: Primary survey attributes and levels

| Level | Age | Initial utility | Initial life expectancy | Final utility | Gain in life expectancy | Patients |
|-------|-----|-----------------|-------------------------|---------------|-------------------------|----------|
| 1 | 10 | .1 | 1 month | .1 | 1 year | 100 |
| 2 | 40 | .5 | 5 years | .5 | 5 years | 2,500 |
| 3 | 70 | .9 | 10 years | .9 | 10 years | 5,000 |

Respondents were told that the patient groups all had some form of cancer, but specific diagnoses were not mentioned (i.e. the alternatives were unlabelled). It was hoped that this additional context would allow all respondents to understand the choice tasks, and their attributes and levels, in a more comparable, consistent manner. To ensure a focus on the attribute levels and not the disease labels, the alternatives were presented simply as Program A and Program B. Although labelled alternatives have the advantage of making hypothetical choice tasks more realistic and concrete, respondents may also use such labels to infer information that was not presented – or intended – as part of the task (de Bekker-Grob, Hol, et al. 2010). At the extreme, respondents may ignore trade-offs between labelled alternatives and make their choices based on their perceptions of the labels alone (Amaya-Amaya et al. 2008).

The experimental design process began with the set of 594 logical scenarios from the pilot design, excluding combinations where the net QALY gain with treatment was negative as well as scenarios where health state and life expectancy were unchanged before and after treatment. The Fedorov algorithm

used to derive the D-efficient design was able to take advantage of the preference weights derived from the pilot survey to estimate the expected choice probability for each possible scenario given the pre-specified value function. This value function was defined as continuous main effects with a two-way interaction between life years gained and final health state. The weights for these parameters were derived from a simple multinomial logit model using the combined DCE and CSPC responses, with CSPC responses transformed to discrete choices on the basis of the alternative to which the majority of the budget was allocated. The results of this model are shown in Appendix 6.1. The incorporation of these weights meant that the algorithm was able to select scenario pairs that balanced scenario utility while also respecting the other design principles, leading to a more statistically efficient optimal fractional factorial design (Huber & Zwerina 1996; Carlsson & Martinsson 2003; Kuhfeld 2010). This was an intentionally simple model, but as Carlsson and Martinsson (2003) noted, D-efficient designs appear to be robust against biased priors, and biased D-efficient designs still lead to more precise parameter estimates than orthogonal designs.

Given this specification, the smallest feasible design was 18 choice sets, and to minimise the burden on individual respondents, it was again divided into 2 blocks of 9 tasks each. Each block also included a test of dominance, or non-satiation, and a repeated task to test preference stability. In the dominance task, two alternatives with identical levels of age, initial health state and number of patients treated, all set to their middle level, were presented to respondents. Final health state and life years gained were also included at their middle levels in one alternative, while the other alternative included them at their highest level. In this choice task, one alternative was unambiguously better in terms of health gain and was intended to test non-satiation in respondents (Miguel et al. 2005; Ryan 2009). This was presented as the first choice in all versions of the design, in hopes that it would be a relatively easy choice that would ease respondents into the elicitations (Carson et al. 1994). In the repeated task, the two alternatives presented in task 5 of each block were reversed and re-presented as task 8. It was felt that the greater incidence of inconsistent preferences in the pilot CSPC relative to the DCE may have been due in part to a longer learning

process with the more complex format. Therefore, the initial choice set was presented as the fifth rather than the third task in order to allow respondents to become more familiar with their preferences in the context of the choice tasks. This task was re-presented as the eighth task to allow respondents some time to forget the original task, yet not so late as to risk significant fatigue effects. As in the pilot survey, these tests of rationality were used only to compare response behaviours in the two questionnaires, and were not used to exclude ‘irrational’ respondents (Lancsar & Louviere 2006).

Including the tests of non-satiation and stability, each block of the experimental design had a total of 11 choice tasks. The test of dominance was always the first task, and the original and repeated tasks to test stability were always presented as the fifth and eighth tasks, respectively, but the order of the other tasks was systematically rotated, resulting in three versions of each design block. To illustrate, choice set 1 in block 1 was the second task in version 1, the ninth task in version 2, and the fourth task in version 3. Fixing the order of the tests of non-satiation and consistency ensured comparability across versions, while varying the order of the other tasks allowed each task to have a roughly equal chance of being seen at the beginning, middle or end of the elicitation, mitigating ordering effects as well as allowing for the identification of possible learning or fatigue effects.

Responses to the repeated task were excluded from the analysis of preferences as they did not contribute new preference information, and although responses to the dominance task were included as a valid expression of preferences, the task was identical in both blocks and therefore contributed only a single degree of freedom. Over the two blocks, counting only one degree of freedom for the common test of dominance, this provided 19 degrees of freedom as understood in the context of conjoint analysis (Hensher et al. 2005).

6.1.2 Data collection

The survey population was drawn from two groups: an age and gender representative sample of the Canadian general public, and a convenience sample of Canadian decision-making agents in oncology, including funding and formulary committee members and oncology professionals. As with the cancer

context of the survey, limiting agents to oncology decision-makers and professional was intended to encourage a common understanding of the different attribute levels, but it was also for pragmatic reasons. There are a relatively large number of oncology decision-making bodies in Canada, so agents in this area may be more familiar with their prioritisation preferences than agents in other disease areas.

The general population sample was drawn from an online survey panel maintained by Research Now™, a market research firm. There is no formal sample size calculation for choice-based stated preference elicitation, but Orme (2006b) offered the following rule-of-thumb:

$$\frac{nta}{c} \geq 500 \quad (6.1)$$

Where n is the number of respondents, t is the number of choice tasks each respondent is presented, a is the number of alternatives per task, and c is the largest number of levels in an attribute. For models that include interactions, c is the product of the number of levels in the largest interaction. Re-arranging equation 6.1 to solve for n , and using the design characteristics described in section 6.1.1, allowing for two-way interactions between attributes and not counting the test of dominance or the repeated task, yielded a minimum sample size of 250 respondents per design block:

$$n \geq \frac{500c}{ta}$$

$$n \geq \frac{500(3 \times 3)}{(9)(2)} \quad (6.2)$$

$$n \geq 250$$

Given two design blocks in each of the two questionnaire formats, this implied a minimum sample of 1000 respondents.

A quota was defined for each combination of sex and 10-year age group to in order to match the Canadian age-sex distribution. Information on respondent income and education was provided by Research Now™, but these characteristics were not included among the quota criteria as this would have substantially complicated recruitment without clearly improving the

representativeness of the sample. Respondents were allocated to the DCE or CSPC questionnaire using a form of sequential balancing, by which each respondent was assigned to the design with the lower number of completed questionnaires for their age-sex subgroup (Borm et al. 2005). This approach ensured that the number and demographic characteristics of respondents would be balanced between the two elicitation formats.

Potential decision-making agents were invited to participate via email and flyers distributed by the pan-Canadian Oncology Drug Review, the Canadian Association of Medical Oncologists, the Canadian Centre for Applied Research in Cancer Control, and provincial cancer authorities including Cancer Care Nova Scotia, Cancer Care Ontario, and Cancer Care British Columbia. All respondents to these agent invitations were allowed to participate in the survey regardless of their age and gender, but they were asked to identify themselves as health system decision makers, including members of decision making committees, program or formulary managers, and health technology assessment practitioners, and/or as oncology professionals. Individuals not self-identifying as one or more of these decision-making groups were categorised as general public. Questionnaires were administered via the internet.

6.1.3 Choice and ratings tasks

As in the pilot elicitation, respondents were asked to imagine themselves as a societal decision maker responsible for allocating a fixed budget between two alternative healthcare programs. They were told that both programs had the same cost, and that the budget was large enough fund one program or the other, but not both. The DCE questionnaire asked respondents to allocate the entire budget to their preferred group, while the CSPC questionnaire asked respondents to allocate budget percentages between the two groups by moving a slider. In each CSPC choice task, the number of patients treated and aggregate QALYs gained changed in proportion to the budget as a respondent moved the slider, and the position of the slider was randomised between each task in order to avoid anchoring effects. Although very few respondents took advantage of the 'no answer' option in the DCE questionnaire, this option was included again in order to encourage questionnaire completion.

The presentation of the choice tasks was changed slightly based on comments from the pilot survey. The table of attributes and levels in each choice task was re-arranged, so that the levels in each alternative were closer together and could be compared more easily. A line was added to the beginning of the instruction following the first task to inform respondents that it had not changed and was the same as in the previous task. Respondents to the pilot survey had complained that they had had to read the full instruction each time to see if anything had changed. A line was also added to note that resources could not be used for research that might improve a patient's condition in the future, as some pilot respondents had asked about this possibility. QALY graphs were included in each choice task to illustrate the magnitude of individual QALY gains with each of the two alternatives. The graphs were similar to those used by Dolan et al. (2008) and Baker et al. (2010), and also illustrated age at disease onset and death, and the patient's initial health state and health state with/after treatment. A limitation of the graphs was that they illustrated individual rather than aggregate QALY gains, but a note was added to each graph to highlight this fact and to encourage respondents to also consider the number of patients treated in their decisions. See Appendix 6.3 for sample DCE and CSPC choice tasks and QALY graphs.

As noted in Chapter 1, the primary survey was conducted in the context of cancer. This was largely for pragmatic reasons, but it was hoped that a defined disease context would encourage respondents to take their choices more seriously, and to understand the levels of the different alternatives – particularly survival and quality gains – in a more comparable manner (Amaya-Amaya et al. 2008; de Bekker-Grob, Hol, et al. 2010). The characteristics of the different patient groups presented in the choice tasks, though, were purely hypothetical and specific diagnoses were not mentioned. The introduction to the questionnaires is shown in Appendix 6.3.

Respondents from the general population sample received rewards for submitting complete – though not necessarily well-considered – questionnaires, and this may have led some respondents to 'click through' the questionnaires without fully considering their answers. Louviere et al. (2000b) argue that such responses appear as random 'statistical noise' rather than a systematic bias, but

to assess the impact of potentially unconsidered responses, individuals who completed the general population surveys in less than one half of the median completion time of each design were flagged as 'fast completers' and preference weights were re-estimated excluding these respondents. Completion times were not available for respondents to the agent invitations, but as there was no reward for completing this version of the survey, there was little reason to expect these respondents would submit random or unconsidered responses. In addition, CSPC respondents who did not move the slider from its initial random position in any of their responses and who were also fast responders were classified as non-informative 'static responders' and excluded from the analyses. Although it is possible that the randomised initial positions exactly matched the static respondents' preferences, the likelihood of such a series of random coincidences seems so small that these exclusions do not appear to be a case of imposing preferences.

Following the choice tasks, respondents were asked to rate the importance of each attribute, including total QALYs gained and distributional concerns, in their choices on a 0 to 8 scale. The 8-point scale reflected the number of factors that respondents were asked to rate, and it was hoped that this would encourage respondents to consider the ratings in terms of relative importance. An actual ranking exercise was not used as it was thought that it would be too challenging for respondents, particularly after having already completed the choice tasks, and that it might discourage respondents from completing the task. Respondents were also asked to rate the difficulty of understanding the tasks, and of answering them, on 5-point scales ranging from extremely easy to extremely difficult, and to indicate "How confident are you that your answers in this survey accurately reflect your preferences for how healthcare resources should be allocated?" on a similar 5-point scale, ranging from very confident to not at all confident.

Attitudes toward healthcare rationing and public participation in healthcare decision making were also elicited. First, respondents were asked to indicate their agreement with the statement "It is impossible for any government or healthcare system to pay for all new medical treatments or technologies, so difficult funding choices will always have to be made" on a 4-point scale ranging

from strongly agree to strongly disagree. No neutral or undecided option was included on the scale in order to force respondents to express agreement or disagreement. Second, they were asked “Who do you think should make the decisions about whether or not different programs should be funded?” Respondents were able to pick one or more of health system decision makers and other experts, doctors and nurses, patients, and citizens or the general public. There was also a text field to allow respondents to enter their own suggestions. Finally, respondents were asked “How comfortable would you be if your preferences were used in determining the allocation of healthcare resources to different programs?” on a 5-point scale ranging from extremely comfortable to extremely uncomfortable. The wording and layout of these ratings tasks is shown in Appendix 6.4.

The questionnaires and the subsequent data analyses were approved by The University of Sheffield Research Ethics Committee, Sheffield UK, and the Capital Health Research Ethics Board, Halifax, Nova Scotia, Canada.

6.2 Primary experimental design

The primary experimental design, with attribute-level combinations for each block and version, is shown in Appendix 6.1. Table 6.2 shows the Pearson correlation coefficients between the parameters included in the value function specified in the experimental design phase, including the life years gained-final health state interaction term, as well as the correlations between the attributes and the block and alternative assignments in the presentation of the choice tasks. Ideally, the attributes should not be correlated with each other or with the block and alternative in which they appear, but the non-orthogonal optimal fractional factorial design that was used here, as well as the exclusion of illogical scenarios, makes some correlation inevitable. Absolute correlations equal to or greater than a moderate threshold of 0.30 (Cohen 1988) are shown in bold.

Table 6.2: Experimental design attribute correlations

| | LYg | Age | U0 | LE0 | U1 | Pats | LYg:U1 | Block | Alt |
|--------|-----|------|-------|-------|-------------|-------|-------------|-------|-------------|
| LYg | --- | 0.10 | 0.01 | -0.04 | -0.03 | 0.17 | 0.66 | 0.07 | 0.07 |
| Age | | --- | -0.07 | 0.13 | 0.14 | 0.05 | 0.15 | -0.07 | 0.50 |
| U0 | | | --- | -0.08 | 0.37 | 0.09 | 0.19 | -0.07 | 0.22 |
| LE0 | | | | --- | 0.08 | -0.03 | 0.02 | 0.00 | -0.14 |
| U1 | | | | | --- | -0.08 | 0.61 | 0.04 | 0.24 |
| Pats | | | | | | --- | 0.09 | 0.00 | -0.27 |
| LYg:U1 | | | | | | | --- | 0.19 | 0.21 |
| Block | | | | | | | | --- | 0.00 |
| Alt | | | | | | | | | --- |

LYg=individual life years gained; Age=Patient age; U0=initial utility; LE0=initial life expectancy; U1=final life expectancy; nPats=total patients treated; LYg:U1= LYg×U1 interaction term; Block=Design block; Alt=Choice alternative. Correlations equal to or greater than a moderate threshold of 0.30 are shown in bold.

Not surprisingly, there were strong positive correlations between the life years gained-final health state interaction term (*LYg:U1*) and its components, *LYg* ($r=0.66$) and *U1* ($r=0.61$). Correlations among the main effects were generally low, although there was a moderate positive correlation ($r=0.37$) between initial health state (*U0*) and final health state (*U1*). This was most likely driven by the exclusion of scenarios with no QALY gains or negative QALY gains from the candidate design, forcing *U1* to be at least equal to *U0* in most scenarios. With respect to the presentation of attribute levels in the choice tasks, there was a moderate correlation between age and alternative ($r=0.50$), suggesting that Alternative B (the right-hand side of the choice task) may have tended to present scenarios with older patients than Alternative A (the left-hand side of the choice task).

6.3 Sample characteristics

Data collection for the general population survey began on 31 January 2012 and continued until the quota of 1,000 respondents was met on 7 February 2012. However, due to a misspecification of the sampling frame, the initial quotas generated a sample that was representative by age and sex independently, rather than jointly. Respondents were representative of the Canadian population by age independent of sex, and representative by sex independent of age, but not representative over age and sex jointly, and younger females and older males were substantially over-represented in the sample. To correct this, the survey

was re-opened on 1 August 2012 but was restricted to participants from the under-represented age-sex groups. Data collection continued until enough respondents in the under-represented age-sex groups were added to allow for the possibility of an age-sex representative sub-sample of at least 1000 respondents. This second phase of the data collection was closed on 3 August 2012. Data collection from the agent invitations began on 30 November 2011 and continued until 22 March 2012. Respondents to the agent invitations were initially allocated between the CSPC and DCE on a 60-40 basis, as in the pilot survey, but this was adjusted to an even 50-50 split when it appeared that responses to the CSPC were outnumbering those to the DCE mid-way through the survey.

The combined agent and general population surveys collected 1,318 completed questionnaires: 656 from the DCE and 662 from the CSPC. The distribution of survey respondents by sex and age group, relative to the Canadian age-sex distribution from the 2011 census (Government of Canada 2012), is shown in Table 6.3.

Table 6.3: Canadian and survey age-sex distributions

| Age group | 2011 Cdn. Census | | Survey | | |
|-----------------|------------------|---------|-----------|-----------|-----------|
| | Males | Females | Males | Females | No answer |
| 18-24 | 6% | 6% | 62 (5%) | 79 (6%) | 0 (0%) |
| 25-34 | 9% | 9% | 109 (8%) | 125 (9%) | 0 (0%) |
| 35-44 | 9% | 8% | 97 (7%) | 213 (16%) | 5 (0%) |
| 45-54 | 10% | 10% | 113 (9%) | 117 (9%) | 3 (0%) |
| 55-64 | 8% | 8% | 89 (7%) | 100 (8%) | 2 (0%) |
| 65-74 | 5% | 5% | 68 (5%) | 31 (2%) | 0 (0%) |
| 75+ | 3% | 5% | 32 (2%) | 71 (5%) | 0 (0%) |
| No answer | - | - | 0 (0%) | 0 (0%) | 2 (0%) |
| Subtotal | 49% | 51% | 570 (43%) | 736 (56%) | 12 (1%) |
| Total | | | | 1,318 | |

As there was little reason to expect that age and gender were the only factors that might influence preferences, it was decided to include the full sample of general population respondents in the analysis, rather than to select a random subset of respondents to generate a representative sample. The full sample was broadly reflective of the Canadian population in terms of their distribution by age and gender, although women in the 35-44 year old age group were substantially over-

represented, while women in the 65-74 year old age group were under-represented. A small number of respondents chose not to answer the demographics questions, but these missing values were not large enough to skew the overall age and gender distribution of the sample.

Among the 1,318 respondents, a total of 101 self-identified as a healthcare decision maker and/or oncology professional. These represented 7.6 percent of all respondents, and the proportion of agents was not significantly different by format: agents represented 6.7 percent of DCE respondents and 8.6 percent of CSPC respondents ($p=0.23$). Among the agents, 39 (39%) identified themselves as health system decision makers, 42 (42%) as oncology professionals, and 20 (20%) as both health system decision makers and oncology professionals.

Overall, the survey respondents appeared to be broadly representative of the Canadian population in terms of the distribution of age, gender, higher education and income. However, there is little reason to believe that such socio-demographic factors are the only observable factors that might influence preferences. Having children, for example, might affect a respondent's preferences for younger age groups, but this characteristic – along with an infinity of other possible confounders – was not accounted for in the sampling frame. In addition, the potential for unobserved heterogeneity or random taste variation among respondents – preferences unrelated to observable characteristics – means that the elicited preferences and attitudes may still be biased or unrepresentative despite the overall socio-demographic representativeness of the sample (Glasgow 2001).

Overall, the survey respondents appeared to be broadly representative of the Canadian population in terms of the distribution of age, gender, higher education and income. However, there is little reason to believe that such socio-demographic factors are the only observable factors that might influence preferences. Having children, for example, might affect a respondent's preferences for younger age groups, but this characteristic – along with an infinity of other possible confounders – was not accounted for in the sampling frame. In addition, the potential for unobserved heterogeneity or random taste variation among respondents – preferences unrelated to observable characteristics – means that the elicited preferences and attitudes may still be biased or

unrepresentative despite the overall socio-demographic representativeness of the sample (Glasgow 2001).

Table 6.4 shows that respondents were well balanced between the two elicitation formats in terms of age and sex. The proportion of general population respondents that had graduated college or university was 35 percent, and this was identical to the 2006 Canadian census population ($p=0.98$) (Statistics Canada 2009), but the median family income category in the sample (\$60,000-64,999) was lower than the 2010 Canadian median family income (\$76,950) (Statistics Canada 2010). There was no significant difference in the proportion of DCE and CSPC general population respondents that graduated college or university (34% vs. 37%, respectively, $p=0.22$), although the median family income category among DCE respondents was slightly lower than among CSPC respondents (\$60,000-65,999 vs. \$65,000-69,999, respectively). Income and education were not collected from respondents to the agent invitations, but it is likely that both would be somewhat higher than in the general population.

Overall, the survey respondents appeared to be broadly representative of the Canadian population in terms of the distribution of age, gender, higher education and income. However, there is little reason to believe that such socio-demographic factors are the only observable factors that might influence preferences. Having children, for example, might affect a respondent's preferences for younger age groups, but this characteristic – along with an infinity of other possible confounders – was not accounted for in the sampling frame. In addition, the potential for unobserved heterogeneity or random taste variation among respondents – preferences unrelated to observable characteristics – means that the elicited preferences and attitudes may still be biased or unrepresentative despite the overall socio-demographic representativeness of the sample (Glasgow 2001).

Table 6.4: Age and sex distribution by questionnaire design

| Age group | DCE | | | CSPC | | |
|--------------|---------|-----------|-----------|---------|-----------|-----------|
| | Male | Female | No answer | Male | Female | No answer |
| 18-24 | 32 (5%) | 39 (6%) | 0 (0%) | 30 (5%) | 40 (6%) | 0 (0%) |
| 25-34 | 61 (9%) | 65 (10%) | 0 (0%) | 48 (7%) | 60 (9%) | 0 (0%) |
| 35-44 | 48 (7%) | 108 (16%) | 1 (0%) | 49 (7%) | 105 (16%) | 4 (1%) |

| | | | | | | |
|------------------|-----------|-----------|--------|-----------|-----------|--------|
| 45-54 | 54 (8%) | 56 (8%) | 1 (0%) | 59 (9%) | 61 (9%) | 2 (0%) |
| 55-64 | 45 (7%) | 45 (7%) | 1 (0%) | 44 (7%) | 55 (8%) | 1 (0%) |
| 65-74 | 35 (5%) | 14 (2%) | 0 (0%) | 33 (5%) | 17 (3%) | 0 (0%) |
| 75+ | 13 (2%) | 36 (5%) | 0 (0%) | 19 (3%) | 35 (5%) | 0 (0%) |
| No answer | 0 (%) | 0 (%) | 2 (0%) | 0 (%) | 0 (%) | 0 (0%) |
| Subtotal | 288 (44%) | 363 (55%) | 5 (1%) | 282 (43%) | 373 (56%) | 7 (1%) |
| Total | 656 (50%) | | | 662 (50%) | | |

6.3.1 Responses by design block and version

Table 6.5 shows that although slightly more respondents were randomised to version 3 of each block, the differences in the overall distribution of respondents between versions were not significant ($p=0.67$).

Table 6.5: Unique respondents by design block and version

| Block | Version | | | All |
|------------|-------------|-------------|-------------|--------------|
| | 1 | 2 | 3 | |
| 1 | 192 (14.6%) | 215 (16.3%) | 231 (17.5%) | 638 (48.4%) |
| 2 | 220 (16.7%) | 224 (17%) | 236 (17.9%) | 680 (51.6%) |
| All | 412 (31.3%) | 439 (33.3%) | 467 (35.4%) | 1,318 (100%) |

$\chi^2=0.8$, $p=0.67$

These counts included respondents to both the DCE and the CSPC questionnaires, but there were more than 300 respondents to each block of each questionnaire format – well above the suggested minimum of 250 respondents per block. Overall the number of respondents to each questionnaire were similar to the samples in a number of recent stated preferences elicitation in healthcare (Green & Gerard 2009; Lancsar et al. 2011; Norman et al. 2013). However, the 101 agents who responded were less than hoped for, and therefore the agent-specific results must be interpreted judiciously.

Table 6.6 shows the same relative distribution of total choices by block and version but accounts for the multiple choices by each individual. With the larger numbers, however, the differences in the overall distribution of responses became statistically significant ($p=0.01$). In general, though, the distribution of responses by version appeared even enough to ensure that each choice task,

excluding the tests of dominance and consistency, was seen at a different stage in the elicitation and had an equal chance of being affected by any learning or fatigue effects.

Table 6.6: Unique choices by design block and version

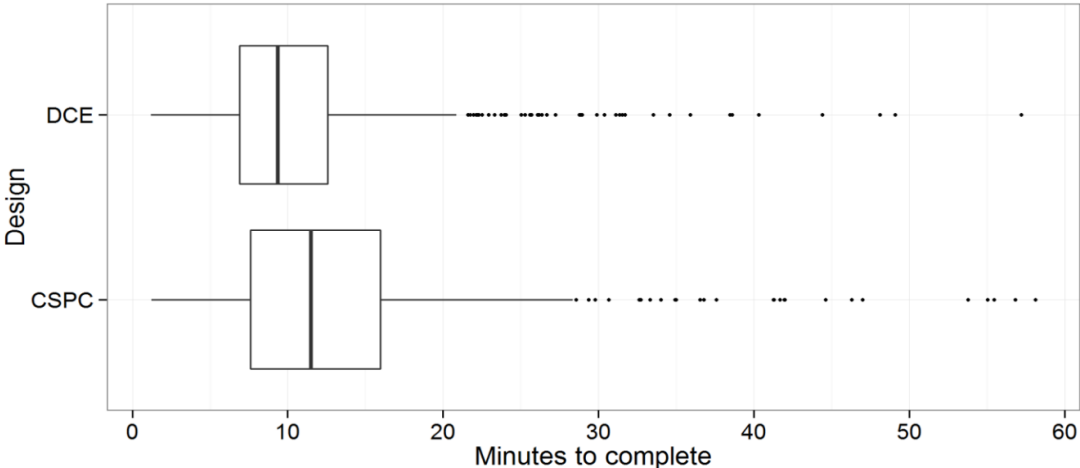
| Block | Version | | | |
|-------|---------------|---------------|---------------|---------------|
| | 1 | 2 | 3 | All |
| 1 | 2,112 (14.6%) | 2,365 (16.3%) | 2,541 (17.5%) | 7,018 (48.4%) |
| 2 | 2,420 (16.7%) | 2,464 (17.0%) | 2,596 (17.9%) | 7,480 (51.6%) |
| All | 4,532 (31.3%) | 4,829 (33.3%) | 5,137 (35.4%) | 14,498 (100%) |

$\chi^2=0.8, p=0.01$

6.3.2 Completion times

The distribution of completion times in the general population sample, excluding times greater than 60 minutes, is illustrated in Figure 6.1. Completion times were not available for respondents to the agent invitations. The median completion time was 9.5 minutes for the DCE and 11.7 minutes for the CSPC. Based on a Mann-Whitney U test, the median completion time among CSPC respondents was significantly greater than among DCE respondents ($p<0.001$).

Figure 6.1: DCE and CSPC completion times, general population respondents only



The boxes show the inter-quartile range (IQR) and contain 50% of the observations from each questionnaire. The heavy vertical lines show the median completion times. The solid horizontal lines (whiskers) extend up to 1.5 times the IQR. Observations outside $\pm 1.5 \times IQR$ are shown as dots (Massart et al., 2005). Completion times greater than 60 minutes are not shown.

The minimum and maximum completion times were 1.2 minutes and 1,244 minutes (20.7 hours), respectively, for the DCE questionnaire, and 1.2 minutes and 4,141 minutes (69.0 hours) for the CSPC questionnaire. Although there were a few extremely high completion times, which were likely respondents who were interrupted and returned to complete the questionnaire at a later time, 98 percent of all the submitted questionnaires were completed in 60 minutes or less.

Among the general population sample, 60 respondents to the DCE (10%) and 75 respondents to the CSPC (13%) had completion times of less than one half the median time for their respective format and were flagged as ‘fast completers.’ These proportions were not significantly different ($p=0.22$). Of the 75 CSPC fast completers, 4 did not move the slider from the initial randomised positions in any choice task. This was taken as overwhelming evidence of inattention in the tasks and their allocation choices were excluded from the choice analyses, although their responses to the attitude questions and the difficulty ratings were included. All respondents to the agent invitations moved the slider in each of their choices.

6.3.3 Respondent attitudes toward rationing

The distribution of respondent attitudes toward the need for healthcare rationing is shown in Table 6.7. The majority of respondents somewhat or strongly agreed with the statement “It is impossible for any government or healthcare system to pay for all new medical treatments or technologies, so difficult funding choices will always have to be made,” although agents were significantly more likely to agree than the general population sample (89% vs. 75%, respectively, $p < 0.01$).

Table 6.7: Rationing attitudes by sample

| Sample | Strongly disagree | Somewhat disagree | Somewhat agree | Strongly agree |
|--------------------|-------------------|-------------------|----------------|----------------|
| General population | 9% | 16% | 46% | 29% |
| Agents | 6% | 5% | 19% | 71% |
| Combined | 9% | 15% | 44% | 32% |
| | 24% | | 76% | |

The proportion of public and agent respondents supporting a decision-making role in healthcare funding decisions for health system decision makers and other experts, doctors, patients and/or citizens is shown in Table 6.8. These categories were not mutually exclusive and respondents could choose none, some, or all groups.

Table 6.8: Proportion of public and agent respondents supporting stakeholder roles in healthcare funding decisions

| Stakeholder group | Public | Agents | p | Adjusted-p | Sig |
|-----------------------------|--------|--------|--------|------------|-----|
| Decision makers and experts | 62% | 85% | <0.001 | <0.001 | *** |
| Doctors and nurses | 71% | 68% | 0.624 | 0.933 | |
| Patients | 46% | 45% | 0.933 | 0.933 | |
| Public | 50% | 56% | 0.251 | 0.754 | |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

Public respondents were most likely to indicate a role for doctors in funding decisions, and least likely to indicate a role for patients. Agents were most likely to indicate a role for health system decision makers and other experts and, like the public respondents, least likely to indicate a role for patients. The only statistically significant difference, though, was in the proportions indicating a role for health system decision makers and other experts, where the public was much less likely than agents to indicate support for their role in funding decisions (adjusted-p < 0.001). Most respondents did not take the opportunity to indicate other groups that should be included in funding decisions, but among those that did, family members was the most frequently mentioned group. This response was somewhat concerning, as it suggests that respondents may have misinterpreted the tasks as patient-level treatment decisions rather than system-level allocation decisions. Other groups mentioned, in no particular order, included scientists, health researchers, ethicists and philosophers, health economists, religious representatives and politicians, including ministers of health. Politicians were also mentioned as a group that should be specifically *excluded* from participating in funding decisions.

Interestingly, support for decision maker and expert involvement in healthcare funding decisions was not universal among self-identified decision-

making agents. One hypothesis for this somewhat counter-intuitive result was that it reflected the differing perspectives of the health system decision makers and physicians that were grouped together as agents. In particular, physicians may have been less likely to support a role for health system decision makers and other experts, preferring to leave such decisions to themselves and other physicians at a more micro-level. This hypothesis was not borne out by the data, though, as physicians were not significantly less likely than health system decision makers to indicate support for decision makers and experts (83% vs. 86%, respectively; adjusted- $p=0.88$). Instead, this lack of universality appears simply to reflect incomplete responses to this question, as agents who did not indicate support for decision makers also showed lower levels of support for all other stakeholder categories.

Finally, the distribution of responses to the question “how comfortable would you be if your preferences were used in determining the allocation of healthcare resources to different programs?” is shown in Table 6.9. Agents were significantly more likely than the public to indicate that they would be somewhat or extremely comfortable having their preferences used for priority setting (65.3% vs. 52.8%, $p=0.02$).

Table 6.9: Proportions by comfort with having their preferences used in priority setting decisions

| Preference comfort | Public | Agents |
|---------------------------------------|--------|--------|
| Extremely comfortable | 10% | 8% |
| Somewhat comfortable | 43% | 58% |
| Neither comfortable nor uncomfortable | 28% | 15% |
| Somewhat uncomfortable | 15% | 16% |
| Extremely uncomfortable | 5% | 3% |

Fisher's Exact Test, $p = 0.018$. Proportions may not sum to 100% due to rounding.

Although the proportion of general population sample who reported that they would be somewhat or extremely comfortable (53%) was very similar to the proportion who were comfortable with giving the public a role in priority setting decisions (50%), Table 6.10 suggests that there was only a slight and statistically insignificant association between a respondents own comfort and their support for a public role in priority setting.

Table 6.10: General public support for a public role in decision making by own preference comfort

| Own preference comfort | Support public role? | |
|-----------------------------------|----------------------|-----|
| | Yes | No |
| Somewhat or extremely comfortable | 52% | 48% |
| Neutral or uncomfortable | 46% | 54% |

Fisher's Exact Test, $p = 0.57$. Proportions may not sum to 100% due to rounding.

6.4 Implications for democratic or Communitarian priority setting

The higher support among the general public for physician involvement in healthcare rationing decisions relative to their support for public or patient involvement appears consistent with an implicit principal-agent relationship in healthcare rationing decisions. The public appeared to prefer physicians to act as their agents in these decisions over health system decision makers and other experts, and this is perhaps not surprising, given the well-established doctor-patient agency relationship in healthcare (Ryan 1994), as well as evidence of a distrust of top-down rationing by experts (Leonard 2012) and of government in general (Edelman 2013). Agents expressed slightly greater support for a public role in priority setting, which is consistent with suggestions that agents may support public involvement as a way of “sharing a bit of the pain” (Coast 2001a), or forcing the public to take ownership of the tough choices (Lomas 1997) involved in making difficult allocation decisions. However, this support was still substantially lower than their support for health system decision-makers and physicians.

The level of support among the general population sample for public involvement in priority setting, although relatively low, appears slightly stronger than results from other work in this area. Lomas (1997) and Coast (2001a) both found that members of the public generally felt that they lacked sufficient knowledge and objectivity to contribute to rationing decisions, and that they would prefer to avoid the responsibility of rationing care. Interestingly, Abelson et al. (1995) found that although only 17 percent of the public felt that *other* members of the public should have a decision-making role in healthcare priority setting, 61 percent were willing to take a *personal* decision-making role. This suggested that the lack of public support for public participation may be less

about reservations over taking responsibility for rationing decisions and more about concerns over the suitability of others for this role. Such a distinction between support for a personal role and a role for the general public at large was not observed in this survey.

Litva et al. (2002) found that the public's willingness to participate in healthcare decision making increased as the identifiability of the patient decreased. Only 21 percent of their study participants were willing to be involved in patient-level allocation decisions, citing reservations about responsibility and denial disutility, but 68 percent were willing to participate in system-level decisions. The level to which respondents to this survey felt they were contributing is not clear, but Litva et al.'s (2002) observation suggests that clearly explaining that preferences would be used to define system objectives and not to discriminate between individual patients may increase support for a public role and the willingness to have one's preferences used in priority setting. Litva et al. also found that participants were more willing to participate if they felt that they would have a real opportunity to influence decisions, rather than just a consultative role. This suggests that although a democratic or Communitarian approach to priority setting may not be embraced by all citizens, a process where all citizens have an genuine opportunity to participate in setting system-level objectives may generate greater citizen support and participation than indirect processes where only a small number of 'representative' citizens express an opinion.

Appendix 6.1: Pilot preference weights used in developing the primary experimental design

| Parameter | Estimate | Std. Error | t-value | Pr(> t) | Sig |
|-----------|----------|------------|---------|----------|-----|
| LYg | 0.251 | 0.027 | 9.318 | <0.001 | *** |
| Age | -0.019 | 0.002 | -12.598 | <0.001 | *** |
| U0 | -0.774 | 0.094 | -8.226 | <0.001 | *** |
| LE | -0.026 | 0.007 | -3.671 | <0.001 | *** |
| U1 | 3.204 | 0.268 | 11.970 | <0.001 | *** |
| nPats | 0.000 | 0.000 | 5.566 | <0.001 | *** |
| U1:LYg | -0.173 | 0.045 | -3.820 | <0.001 | *** |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+'

LYg=Life years gained; U0=Initial utility; LE=Initial life expectancy; U1=Final utility; nPats=Number of patients treated

Appendix 6.2: Primary experimental design, by block and version

| Block | Version | Choice | Set | Alternative A (left-hand side) | | | | | | Alternative B (right-hand side) | | | | | | Tests | | |
|-------|---------|--------|-----|--------------------------------|-------------|-------------------------|-----------|------------|------------------|---------------------------------|-------------|-------------|-------------------------|-----------|------------|-------|------------------|--------------------|
| | | | | Patient age | Initial QoL | Initial life expectancy | Final QoL | Lys gained | Patients treated | Aggregate QALYs | Patient age | Initial QoL | Initial life expectancy | Final QoL | Lys gained | | Patients treated | Aggregate QALYs |
| 1 | 1 | 1 | 10 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | 40 | 1 | 0.083 | 5 | 5 | 2,500 | 6,333 | Test non-satiation |
| 1 | 1 | 2 | 1 | 10 | 1 | 10 | 1 | 5 | 2,500 | 1,250 | 40 | 9 | 0.083 | 5 | 10 | 5,000 | 24,834 | |
| 1 | 1 | 3 | 2 | 10 | 5 | 10 | 5 | 1 | 2,500 | 1,250 | 40 | 1 | 5 | 1 | 100 | 10 | | |
| 1 | 1 | 4 | 3 | 40 | 9 | 0.083 | 1 | 5 | 2,500 | 1,084 | 70 | 5 | 10 | 9 | 1 | 100 | 490 | |
| 1 | 1 | 5 | 7 | 10 | 5 | 5 | 9 | 10 | 100 | 1,100 | 70 | 1 | 0.083 | 9 | 5 | 5,000 | 22,832 | Original task |
| 1 | 1 | 6 | 4 | 10 | 5 | 0.083 | 1 | 10 | 5,000 | 4,834 | 70 | 9 | 10 | 5 | 10 | 2,500 | 2,500 | |
| 1 | 1 | 7 | 5 | 70 | 1 | 5 | 5 | 5 | 2,500 | 11,250 | 40 | 9 | 10 | 9 | 1 | 100 | 90 | |
| 1 | 1 | 8 | 77 | 70 | 1 | 0.083 | 9 | 5 | 5,000 | 22,832 | 10 | 5 | 5 | 9 | 10 | 100 | 1,100 | Repeated task |
| 1 | 1 | 9 | 6 | 40 | 1 | 5 | 5 | 1 | 5,000 | 12,500 | 10 | 5 | 0.083 | 1 | 10 | 2,500 | 2,417 | |
| 1 | 1 | 10 | 8 | 40 | 1 | 10 | 1 | 10 | 5,000 | 5,000 | 70 | 5 | 0.083 | 5 | 5 | 100 | 250 | |
| 1 | 1 | 11 | 9 | 10 | 9 | 0.083 | 9 | 1 | 5,000 | 4,500 | 40 | 5 | 10 | 9 | 5 | 100 | 850 | |
| 1 | 2 | 1 | 10 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | 40 | 1 | 0.083 | 5 | 5 | 2,500 | 6,333 | Test non-satiation |
| 1 | 2 | 2 | 4 | 10 | 5 | 0.083 | 1 | 10 | 5,000 | 4,834 | 70 | 9 | 10 | 5 | 10 | 2,500 | 2,500 | |
| 1 | 2 | 3 | 5 | 70 | 1 | 5 | 5 | 5 | 2,500 | 11,250 | 40 | 9 | 10 | 9 | 1 | 100 | 90 | |
| 1 | 2 | 4 | 6 | 40 | 1 | 5 | 5 | 1 | 5,000 | 12,500 | 10 | 5 | 0.083 | 1 | 10 | 2,500 | 2,417 | |
| 1 | 2 | 5 | 7 | 10 | 5 | 5 | 9 | 10 | 100 | 1,100 | 70 | 1 | 0.083 | 9 | 5 | 5,000 | 22,832 | Original task |
| 1 | 2 | 6 | 8 | 40 | 1 | 10 | 1 | 10 | 5,000 | 5,000 | 70 | 5 | 0.083 | 5 | 5 | 100 | 250 | |
| 1 | 2 | 7 | 9 | 10 | 9 | 0.083 | 9 | 1 | 5,000 | 4,500 | 40 | 5 | 10 | 9 | 5 | 100 | 850 | |
| 1 | 2 | 8 | 77 | 70 | 1 | 0.083 | 9 | 5 | 5,000 | 22,832 | 10 | 5 | 5 | 9 | 10 | 100 | 1,100 | Repeated task |
| 1 | 2 | 9 | 1 | 10 | 1 | 10 | 1 | 5 | 2,500 | 1,250 | 40 | 9 | 0.083 | 5 | 10 | 5,000 | 24,834 | |
| 1 | 2 | 10 | 2 | 10 | 5 | 10 | 5 | 1 | 2,500 | 1,250 | 40 | 1 | 5 | 1 | 100 | 10 | | |
| 1 | 2 | 11 | 3 | 40 | 9 | 0.083 | 1 | 5 | 2,500 | 1,084 | 70 | 5 | 10 | 9 | 1 | 100 | 490 | |

| | | | | | | | | | | | | | | | | | | |
|---|---|----|----|----|---|-------|---|----|-------|--------|----|---|-------|---|----|-------|--------|--------------------|
| 1 | 3 | 1 | 10 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | 40 | 1 | 0.083 | 5 | 5 | 2,500 | 6,333 | Test non-satiation |
| 1 | 3 | 2 | 8 | 40 | 1 | 10 | 1 | 10 | 5,000 | 5,000 | 70 | 5 | 0.083 | 5 | 5 | 100 | 250 | |
| 1 | 3 | 3 | 9 | 10 | 9 | 0.083 | 9 | 1 | 5,000 | 4,500 | 40 | 5 | 10 | 9 | 5 | 100 | 850 | |
| 1 | 3 | 4 | 1 | 10 | 1 | 10 | 1 | 5 | 2,500 | 1,250 | 40 | 9 | 0.083 | 5 | 10 | 5,000 | 24,834 | Original task |
| 1 | 3 | 5 | 7 | 10 | 5 | 5 | 9 | 10 | 100 | 1,100 | 70 | 1 | 0.083 | 9 | 5 | 5,000 | 22,832 | |
| 1 | 3 | 6 | 2 | 10 | 5 | 10 | 5 | 1 | 2,500 | 1,250 | 40 | 1 | 5 | 1 | 1 | 100 | 10 | |
| 1 | 3 | 7 | 3 | 40 | 9 | 0.083 | 1 | 5 | 2,500 | 1,084 | 70 | 5 | 10 | 9 | 1 | 100 | 490 | |
| 1 | 3 | 8 | 77 | 70 | 1 | 0.083 | 9 | 5 | 5,000 | 22,832 | 10 | 5 | 5 | 9 | 10 | 100 | 1,100 | Repeated task |
| 1 | 3 | 9 | 4 | 10 | 5 | 0.083 | 1 | 10 | 5,000 | 4,834 | 70 | 9 | 10 | 5 | 10 | 2,500 | 2,500 | |
| 1 | 3 | 10 | 5 | 70 | 1 | 5 | 5 | 5 | 2,500 | 11,250 | 40 | 9 | 10 | 9 | 1 | 100 | 90 | |
| 1 | 3 | 11 | 6 | 40 | 1 | 5 | 5 | 1 | 5,000 | 12,500 | 10 | 5 | 0.083 | 1 | 10 | 2,500 | 2,417 | |
| 2 | 1 | 1 | 10 | 40 | 1 | 0.083 | 5 | 5 | 2,500 | 6,333 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | Test non-satiation |
| 2 | 1 | 2 | 1 | 10 | 5 | 5 | 5 | 10 | 5,000 | 25,000 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | |
| 2 | 1 | 3 | 2 | 40 | 5 | 10 | 5 | 10 | 2,500 | 12,500 | 10 | 9 | 5 | 9 | 5 | 100 | 450 | |
| 2 | 1 | 4 | 3 | 10 | 5 | 0.083 | 1 | 1 | 5,000 | 334 | 70 | 1 | 5 | 1 | 10 | 100 | 100 | Original task |
| 2 | 1 | 5 | 8 | 40 | 5 | 5 | 9 | 1 | 2,500 | 7,250 | 70 | 9 | 10 | 9 | 10 | 5,000 | 45,000 | |
| 2 | 1 | 6 | 4 | 70 | 1 | 10 | 1 | 5 | 5,000 | 2,500 | 40 | 9 | 0.083 | 5 | 1 | 2,500 | 1,167 | |
| 2 | 1 | 7 | 5 | 10 | 1 | 10 | 5 | 5 | 100 | 650 | 70 | 5 | 5 | 9 | 10 | 5,000 | 55,000 | Repeated task |
| 2 | 1 | 8 | 88 | 70 | 9 | 10 | 9 | 10 | 5,000 | 45,000 | 40 | 5 | 5 | 9 | 1 | 2,500 | 7,250 | |
| 2 | 1 | 9 | 6 | 10 | 1 | 0.083 | 5 | 10 | 100 | 503 | 40 | 5 | 10 | 5 | 5 | 5,000 | 12,500 | |
| 2 | 1 | 10 | 7 | 10 | 1 | 10 | 1 | 1 | 2,500 | 250 | 40 | 5 | 0.083 | 1 | 5 | 100 | 47 | |
| 2 | 1 | 11 | 9 | 10 | 9 | 5 | 9 | 5 | 2,500 | 11,250 | 40 | 1 | 0.083 | 5 | 1 | 100 | 53 | |
| 2 | 2 | 1 | 10 | 40 | 1 | 0.083 | 5 | 5 | 2,500 | 6,333 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | Test non-satiation |
| 2 | 2 | 2 | 4 | 70 | 1 | 10 | 1 | 5 | 5,000 | 2,500 | 40 | 9 | 0.083 | 5 | 1 | 2,500 | 1,167 | |
| 2 | 2 | 3 | 5 | 10 | 1 | 10 | 5 | 5 | 100 | 650 | 70 | 5 | 5 | 9 | 10 | 5,000 | 55,000 | Original task |
| 2 | 2 | 4 | 6 | 10 | 1 | 0.083 | 5 | 10 | 100 | 503 | 40 | 5 | 10 | 5 | 5 | 5,000 | 12,500 | |
| 2 | 2 | 5 | 8 | 40 | 5 | 5 | 9 | 1 | 2,500 | 7,250 | 70 | 9 | 10 | 9 | 10 | 5,000 | 45,000 | |
| 2 | 2 | 6 | 7 | 10 | 1 | 10 | 1 | 1 | 2,500 | 250 | 40 | 5 | 0.083 | 1 | 5 | 100 | 47 | |
| 2 | 2 | 7 | 9 | 10 | 9 | 5 | 9 | 5 | 2,500 | 11,250 | 40 | 1 | 0.083 | 5 | 1 | 100 | 53 | |
| 2 | 2 | 8 | 88 | 70 | 9 | 10 | 9 | 10 | 5,000 | 45,000 | 40 | 5 | 5 | 9 | 1 | 2,500 | 7,250 | Repeated task |

| | | | | | | | | | | | | | | | | | | | |
|---|---|----|----|----|----|---|-------|---|----|-------|--------|----|---|-------|---|----|-------|--------|--------------------|
| 2 | 2 | 2 | 9 | 1 | 10 | 5 | 5 | 5 | 10 | 5,000 | 25,000 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | |
| 2 | 2 | 2 | 10 | 2 | 40 | 5 | 10 | 5 | 10 | 2,500 | 12,500 | 10 | 9 | 5 | 9 | 5 | 100 | 450 | |
| 2 | 2 | 2 | 11 | 3 | 10 | 5 | 0.083 | 1 | 1 | 5,000 | 334 | 70 | 1 | 5 | 1 | 10 | 100 | 100 | |
| 2 | 3 | 1 | 1 | 10 | 40 | 1 | 0.083 | 5 | 5 | 2,500 | 6,333 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | |
| 2 | 3 | 2 | 7 | 10 | 10 | 1 | 10 | 1 | 1 | 2,500 | 250 | 40 | 5 | 0.083 | 1 | 5 | 100 | 47 | |
| 2 | 3 | 3 | 9 | 10 | 10 | 9 | 5 | 9 | 5 | 2,500 | 11,250 | 40 | 1 | 0.083 | 5 | 1 | 100 | 53 | |
| 2 | 3 | 4 | 1 | 10 | 10 | 5 | 5 | 5 | 10 | 5,000 | 25,000 | 40 | 1 | 0.083 | 9 | 10 | 2,500 | 22,666 | |
| 2 | 3 | 5 | 8 | 40 | 40 | 5 | 5 | 9 | 1 | 2,500 | 7,250 | 70 | 9 | 10 | 9 | 10 | 5,000 | 45,000 | |
| 2 | 3 | 6 | 2 | 40 | 40 | 5 | 10 | 5 | 10 | 2,500 | 12,500 | 10 | 9 | 5 | 9 | 5 | 100 | 450 | |
| 2 | 3 | 7 | 3 | 10 | 10 | 5 | 0.083 | 1 | 1 | 5,000 | 334 | 70 | 1 | 5 | 1 | 10 | 100 | 100 | |
| 2 | 3 | 8 | 88 | 70 | 70 | 9 | 10 | 9 | 10 | 5,000 | 45,000 | 40 | 5 | 5 | 9 | 1 | 2,500 | 7,250 | |
| 2 | 3 | 9 | 4 | 70 | 70 | 1 | 10 | 1 | 5 | 5,000 | 2,500 | 40 | 9 | 0.083 | 5 | 1 | 2,500 | 1,167 | |
| 2 | 3 | 10 | 5 | 10 | 10 | 1 | 10 | 5 | 5 | 100 | 650 | 70 | 5 | 5 | 9 | 10 | 5,000 | 55,000 | |
| 2 | 3 | 11 | 6 | 10 | 10 | 1 | 0.083 | 5 | 10 | 100 | 503 | 40 | 5 | 10 | 5 | 5 | 5,000 | 12,500 | |
| | | | | | | | | | | | | | | | | | | | Test non-satiation |
| | | | | | | | | | | | | | | | | | | | Original task |
| | | | | | | | | | | | | | | | | | | | Repeated task |

Appendix 6.3: Sample choice tasks

Introduction


This survey will ask you to **imagine you are responsible for deciding how a healthcare budget should be divided between different groups of cancer patients.**

- The survey is designed to measure your personal preferences. There are no right or wrong answers.
- The type of cancer in each group is not specified. You should make your decision based on the information presented, not the type of cancer you think they might have.
- As you consider your answers, remember that it is possible that you or someone in your family could be part of one of these groups, now or in the future.
- The questions can be quite challenging, and even uncomfortable, but **the results could help improve how money is spent on our healthcare in the future, so your opinions are very important.**

Thank you for your time and attention.

Sample DCE task


| Attributes | Program A | Program B |
|---|-------------------------|-------------------------------|
| Average age of patients | 70 years old | 40 years old |
| Quality-of-life without/before treatment | 9 out of 10 | 5 out of 10 |
| Quality-of-life with treatment | 9 out of 10 [No change] | 9 out of 10 [4 levels higher] |
| Life expectancy without/before treatment | 10 years | 5 years |
| Gain in life expectancy with treatment | 10 additional years | 1 additional year |
| Number of patients that could be treated | 5,000 | 2,500 |
| Total quality-adjusted life years gained with treatment | 45,000 | 7,250 |


 [See this information in a graph](#)

No answer
 I prefer Program A
 I prefer Program B

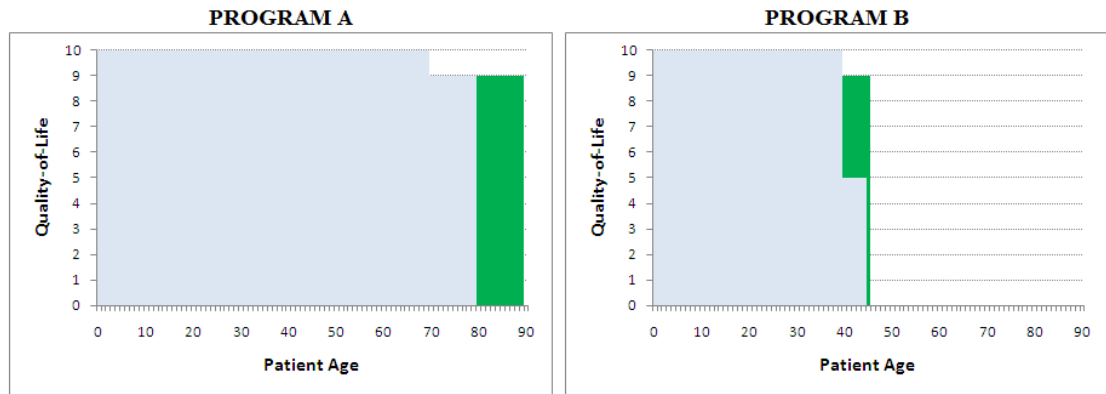
Sample CSPC task

| | Program A | Program B |
|---|-------------------------|-------------------------------|
| Average age of patients | 70 years old | 40 years old |
| Quality-of-life without/before treatment | 9 out of 10 | 5 out of 10 |
| Quality-of-life with treatment | 9 out of 10 [No change] | 9 out of 10 [4 levels higher] |
| Life expectancy without/before treatment | 10 years | 5 years |
| Gain in life expectancy with treatment | 10 additional years | 1 additional year |
| Number of patients that could be treated | 3,600 | 700 |
| Total quality-adjusted life years gained with treatment | 32,400 | 2,030 |

 [See this information in a graph](#)

Percent of budget to Program A **Use this slider to divide the budget between Program A and B** Percent of budget to Program B
 72%  28%

Sample QALY graph



These graphs show quality-of-life by age for patients that could benefit from Program A or Program B. The higher the area, the higher the quality-life. As the graph moves toward the right, the patients are getting older. The light blue area shows length and quality-of-life if patients *do not* receive treatment. In the graphs, patients are in perfect health until they develop cancer, at which point their quality-of-life drops. Once they develop cancer, patients may die almost immediately, or continue to live for some time at a lower quality-of-life. **The green area shows the quality-of-life and years-of-life a patient would gain with treatment.** Because cancer treatment sometimes involves harsh chemotherapy or radiation, it is possible that treatment could *reduce* a patient's quality-of-life in order to extend the length of their life. In this case, the green treatment area will overlap the light blue area as quality-of-life is lower than it would have been without treatment, even though the length-of-life is longer.

Note that these graphs compare individual patient profiles; they DO NOT show the total number of patients treated or total quality-adjusted life years gained with each program.

Appendix 6.4: Rating tasks

1. Do you agree or disagree with the statement, "It is impossible for any government or healthcare system to pay for all new medical treatments or technologies, so difficult funding choices will always have to be made"?

No answer Strongly agree Somewhat agree Somewhat disagree Strongly disagree

2. How **comfortable** would you be if your preferences were used in determining the allocation of healthcare resources to different programs?

No answer Extremely comfortable Somewhat comfortable Neither comfortable nor uncomfortable Somewhat uncomfortable Extremely uncomfortable

3. How easy/difficult did you find it to **understand** the tasks?

No answer Extremely easy Somewhat easy Neither easy nor difficult Somewhat difficult Extremely difficult

4. How easy/difficult did you find it to **answer** the tasks?

No answer Extremely easy Somewhat easy Neither easy nor difficult Somewhat difficult Extremely difficult

5. How **confident** are you that your answers in this survey accurately reflect your preferences for how healthcare resources should be allocated?

No answer Very confident Somewhat confident Undecided/Unclear Not very confident Not at all confident

Chapter 7: Comparison of the DCE and CSPC formats

The pilot survey, discussed in Chapter 5, suggested that the DCE had relatively better completion rates, lower difficulty ratings, and greater preference consistency, suggesting a greater ‘respondent efficiency.’ This had to be weighed against the ability of the CSPC to elicit specific preferences for the distribution of resources and/or outcomes between groups, and to avoid discomfoting extreme distributions by allowing respondents to allocate some resources to the less preferred group. However, the non-orthogonal application of the pilot experimental design meant that respondents to the DCE and CSPC questionnaires did not see the same choice sets. As a result, it was not possible to say with certainty that the observed differences between the two pilot questionnaires were due to the elicitation methods themselves, and not simply due to differences in the choice sets presented to respondents. Therefore, it was of interest to re-compare the two formats on the basis of the larger sample and more appropriate application of the experimental design of the primary elicitation.

It was also useful to compare the response behaviours of the two formats before estimating respondent preferences for the different attributes included in the DCE and CSPC questionnaires. If an excessive proportion of respondents to one questionnaire or the other reported difficulty understanding or answering the tasks, or adopted a non-compensatory decision making strategy, it may call into question the validity of the elicited preferences. The inclusion of respondents with dominant or non-compensatory preferences generally presupposes that they represent a relatively small proportion of all respondents, but if a majority of

respondents adopted such a strategy, it would imply that most attribute levels had no impact on choices, and regression coefficients and rates of substitution would have no meaningful interpretation (Scott 2002). Similarly, high rates of difficulty or simplifying non-compensatory decision-making may indicate an overly complex or confusing elicitation format that may compromise the collection of meaningful preference data (DeShazo & Fermo 2002).

The different dimensions of the comparison of the DCE and CSPC response behaviours, including completion rates, difficulty ratings, tests of non-satiation and preference stability, learning and fatigue effects, dominant preferences, and QALY maximising behaviours, are discussed in section 7.1, with the results of these comparisons presented in section 7.2. Finally, the implications of this comparison for identifying a preferred format for the elicitation of societal preferences over the allocation of healthcare resources are discussed in section 7.3.

7.1 DCE-CSPC comparisons

The response behaviours of the DCE and CSPC respondents were compared on a number of dimensions. These included completion rates, difficulty and confidence ratings, preference stability, and the incidence of dominant preferences, as in the pilot survey, but the primary survey added comparisons of completion times, learning and fatigue effects, and a test of non-satiation or dominance. It is important to note that for all these comparisons each respondent only saw a single questionnaire and never compares the two formats directly. All statistical tests were conducted using R 2.15.3 (R Core Team 2013), and a significance threshold of 0.05 was adopted.

7.1.1 Completion rates

Completion rates for the two formats were compared using a two-sample Z-test of completed questionnaires as a proportion of questionnaires begun. In the general population survey, the assigned questionnaire format, completion status, and (if applicable) completion time, were linked to each individual respondent. Individuals who quit a questionnaire but returned at a later time

could pick up where they left off and submit a completed questionnaire. This would be counted as one survey begun and submitted (a 100% completion rate). Owing to differences in the sample and the survey software, however, responses from the agent invitations were entirely anonymous, and it was not possible to track individuals who may have quit one questionnaire but returned later to complete another. Such a case would be counted as two surveys begun, with one drop-out and one completion (a 50% completion rate). To the extent that respondents to the agent invitations began more than one questionnaire, completion rates in this group will be correspondingly understated.

For the purposes of this comparison, the agent sample included all respondents to the agent invitations, regardless of whether they self-identified as a decision-making agent. As it was not possible to know the status of individuals who started but did not submit a questionnaire, it was not possible to calculate an agent-specific denominator or completion rate. For all other analyses, which only included completed questionnaires, non-agent respondents to the agent invitations were included among the general public sample.

The proportion of ‘no answers’ among DCE responses was also reported, including a comparison of the proportion of no answers between agents and the general public. The trend in the proportion of no answers by task sequence was tested using linear regression. The proportion of no answers in the repeated task was also compared to the proportion across all other tasks combined. The choice set used as the repeated task in each design block was the set with closest utility balance, and there is evidence to suggest that greater utility balance tends to make discrete choice tasks more complex (DeShazo & Fermo 2002). It was of interest, therefore, to test whether this relatively greater complexity may have led respondents to opt out of these tasks at a greater rate than the other tasks.

7.1.2 Respondent-rated difficulty and confidence

The difficulty of the two formats was compared in terms of the proportions of DCE and CSPC respondents who rated the questionnaires as ‘somewhat difficult’ or ‘extremely difficult’ to understand or to answer in the follow-up rating tasks. As the pilot survey suggested that decision makers were more likely to rate the tasks as difficult, this comparison was stratified by agent

status. The questionnaires were also compared in terms of the proportion of respondents who were 'somewhat confident' or 'extremely confident' that their responses to the choice tasks accurately reflected their preferences. Proportions were compared using a two-sample Z-test.

7.1.3 Tests of non-satiation and stability

Each block of the experimental design included a test of each respondent's consistency with the axiom of non-satiation, as well as a repeated task to test preference stability. Stated preference elicitations have generally held that 'rational' respondents should prefer the dominant alternative (Miguel et al., 2005; Ryan, 2009). CSPC allocations were transformed to discrete choices on the basis of which alternative was allocated the majority of the budget, and 50-50 allocations were taken as not prioritising the dominant alternative. The proportion of respondents in each questionnaire demonstrating non-satiation was compared using a two-sample Z-test.

Learning from the pilot survey, where high stability in the repeated DCE task may have been aided by an extreme imbalance in utilities between the two alternatives, the repeated task used in each design block was the choice set with the closest expected utility balance. Again, CSPC budget allocations were transformed to discrete choices on the basis of the alternative to which the majority of resources were allocated. Equal 50-50 allocations were allowed, but the allocations had to be equal in both tasks in order to be classified as consistent. As an equal budget allocation was treated as a distinct choice from Program A or B, a stricter definition of stability was effectively applied to individuals choosing this specific allocation relative to individuals who allocated a majority of the budget to one alternative or the other, as there was only one specific budget allocation that would be accepted as stable.

The proportion of respondents with stable responses to the repeated task was compared using a two-sample Z-test. To assess the impact of the stricter consistency criterion for individuals with an equal budget allocation, the statistical significance of the mean difference in the repeated budget allocations of individuals with at least one 50-50 allocation was tested using a one-sample t-test.

7.1.4 Learning and fatigue effects

Experimental evidence that suggests that there are simultaneous processes of learning and fatigue that may affect choices as respondents progress through a stated preference elicitation (Bech et al., 2011; Johnson and Desvousges, 1997; Maddala et al., 2003). To allow for the identification of such effects, the 11 choice sets in each block of the experimental design were divided into 3 segments, which were systematically rotated to create 3 versions of each block (see Appendix 6.2 for the task sequence within each block and version). The positions of test of non-satiation and the original and repeated tasks to test preference stability were fixed across all versions, but the position of the remaining tasks rotated by version. The order of the tasks in each version is shown in Appendix 6.1.

To test for learning or fatigue effects, a series of one-way analyses of variance (ANOVAs) were conducted for each choice set and design block to test for a statistically significant difference by task sequence in the proportion of DCE respondents that preferred alternative B, or in the mean CSPC budget allocation to alternative B. If a particular choice set was associated with a significant difference, a post hoc test was used to identify which version was the outlier. Preferences that were significantly and systematically different in choice sets seen earlier relative to other versions would suggest a learning effect. Conversely, a significant difference for sets seen later relative to other versions would suggest a fatigue effect. The analyses of variance were performed using the car package in R 2.15.3 (Fox & Weisberg 2011).

7.1.5 Dominant preferences

What appears to be a dominant or non-compensatory preference for a particular attribute may simply be a reluctance to trade over the range of levels presented in the experimental design, or an indication that the attributes included in the experimental design are unimportant to some respondents (Lancsar & Louviere 2006). A fractional factorial experimental design also complicates the interpretation of non-compensatory preferences, as it is not possible to say with certainty that observed instances of non-compensatory decision-making would persist across all possible scenarios (Lancsar & Louviere 2006; Scott 2002). For

these reasons, respondents with a dominant preference are generally included in the analysis of stated preference elicitation. However, this presupposes that such respondents represent a relatively small proportion of all respondents. If all respondents adopted non-compensatory decision making strategies, it would imply that attribute levels had no impact on choices, and would invalidate the stated preference elicitation. Therefore, the incidence of dominant preferences in the two elicitation formats – including a dominant preference for QALY maximisation – was assessed before proceeding with the choice analysis.

As in the pilot survey, a respondent was considered to have a dominant preference for a particular attribute if they always chose or prioritised the alternative with the highest or lowest level of that attribute, regardless of the levels of the other attributes (Scott 2002). To test for dominant preferences, a set of seven flags was created for each alternative in each choice task: age, initial utility, initial life expectancy, final utility, life years gained, (potential) number of patients treated and (potential) aggregate QALYs gained. Each flag indicated whether or not the alternative presented the higher level of a particular attribute. For example, the alternative that included the older patients was flagged as 'highest' in the age attribute, and the flag for the corresponding alternative was set to zero.

The correlation between each individual's choice and attribute flags was taken as a measure of that attribute's impact on their choices. A respondent who, for example, invariably chose the alternative with the youngest patients would have a correlation coefficient of -1.0 between choice flag and the age flag (perfect choice-attribute correlation). Correlation coefficients close to zero would indicate the attribute had relatively little direct impact on their choices. This was slightly different than the approach taken in the pilot survey, where the flags indicated whether the attribute presented the 'more preferred' level based on expectations from the empirical ethics review. In this revised approach, no effort was made to anticipate the preferred level of the attribute, and the sign on the correlation coefficient indicated which end of the attribute scale was preferred (if either): a negative coefficient indicated a preference for the lower level, and a positive coefficient indicated a preference for the higher level. Note, though, that this still only holds where individual preferences are monotonically

increasing or decreasing over the attribute, as was assumed here. CSPC responses were transformed to discrete choices on the basis of which program was allocated the majority of the budget, and the attribute flags were set based on the potential number of patients treated and QALYs gained if 100% of the budget were allocated to that alternative. For the purposes of this analysis, CSPC alternatives that received a 50% budget allocation were flagged as 'not chosen' (i.e. both alternatives in a given task were assigned a choice flag of zero) as neither alternative was prioritised. Kendall's tau (Herve 2007) was used as the measure of correlation, which was estimated using the ltm package (Rizopoulos 2006).

Responses to the test of non-satiation were excluded from the analysis, as the common levels of age, initial health state, initial life expectancy and patients treated in this task meant that both alternatives in the task would be flagged as non-dominant over these attributes. Regardless of which alternative a respondent chose, they would be flagged as choosing the non-dominant alternative, and by definition could not hold a dominant preference for those attributes, even if all their other choices were based on the level of one of those attributes.

To confirm the identification of a possible dominant preference, each respondent's self-rated attribute importance scores were converted to rankings and compared to their choice-attribute correlation coefficients. Individuals with a perfect choice-attribute correlation who also rated that attribute as most important were considered to have a confirmed dominant preference for that attribute. As a respondent could give more than one attribute the same rating, more than one attribute could be ranked as most important, but dominant preferences were considered to be confirmed even in the case of ties with other attributes. The proportion of dominant preferences was compared across the two formats using a two-sample Z-test.

7.1.6 QALY maximisation

The results from the pilot survey suggested that respondents did not, in general, adopt a strict QALY maximising decision rule. Only one respondent prioritised the QALY maximising alternative with every choice, and on average,

DCE and CSPC respondents chose the QALY maximising well less than the 10 times out of 10 that would seem to be required by a strict QALY maximising approach. As noted, however, respondents to the two formats did not see the same choice sets, so it was of interest to test for differences in the number of QALY maximising choices by format when all respondents saw the same choice sets. The mean number of QALY maximising alternatives chosen by each respondent was compared across the formats using a two-sample t-test. CSPC respondents were considered to have prioritised the QALY maximising alternative if that program received the majority of the budget allocation. An equal budget allocation between the two programs was counted as a non-maximising choice. To test whether agents, who may have been more familiar with QALYs and the principles of QALY maximisation than the general public, were significantly more likely to adopt a QALY maximising decision rule, the number of QALY maximising choices was also compared across respondent subgroups. Finally, to test whether the QALY graphs (see Appendix 6.3), which illustrated QALY gains at the individual rather than the aggregate level, may have tended to encourage respondents to focus on individual over aggregate QALY gains, the number of choices that maximised QALYs at the individual level were compared with the number of choices maximising QALYs at the aggregate level.

Note that in each choice task, one of the two alternatives was always the QALY maximising alternative. Therefore, respondents had a 50 percent probability of prioritising the QALY maximising alternative by chance alone, disregarding for now equal CSPC budget allocations. Monte Carlo simulation was used to test whether the observed proportion of QALY maximising choices was significantly different than what might be expected by chance alone, similar to an approach used by Diederich, Swait and Wirsik (2012). Each respondent's vector of observed choices was replaced by a vector of random choices, and the difference in the number of QALY maximising choices between the observed and random vectors calculated. This process was repeated 1000 times and sorted by ascending difference. The differences at the 2.5th and 97.5th percentiles, for agents and for the general population sample, were taken as estimates of the 95 percent confidence intervals of the mean difference between the proportion of

observed QALY maximising choices and what would be expected by chance alone. Statistically significant intervals greater than zero were taken to indicate support for the principles of QALY maximisation, while significantly negative intervals were taken to indicate opposition to this rule. Intervals that crossed zero were taken to indicate no statistical preference for aggregate QALYs.

Finally, empirical evidence (Payne et al. 1992; Ryan 2009; Slovic 1995), as well as anecdotal evidence from the pilot survey, suggested that respondent may construct preferences as they progress through a stated preference elicitation. Therefore, it was also of interest to test whether a respondent's tendency to choose the QALY maximising alternative changed over the task sequence in either format. There were two competing hypotheses: one hypothesis was that respondents would become *more* likely to prioritise the QALY maximising alternative as they became more familiar with the concept of QALYs as a measure of aggregate health gain. The other hypothesis was that respondents would use aggregate QALY gains as a simplifying heuristic in the early tasks, but become *less* likely to prioritise on the basis of QALY gains as they became more familiar with the trade-offs and levels in the choice tasks.

A probit model was used to estimate the change in the probability of a respondent prioritising the QALY maximising alternative by task sequence and questionnaire format. The specific attribute levels in each choice set were disregarded; only the position of the choice in the overall task sequence was considered. The test of non-satiation, presented as task 1 in all questionnaire versions, was excluded from the model as it was felt that the fixed position of the task, as well as the dominance of one alternative, might make it an outlier in the overall trend. Predicted choice probabilities were derived from the probit coefficients using the effects package (Fox 2003), and the standard errors were adjusted for clustering using the lmtest package (Zeileis & Hothorn 2002).

7.2 DCE and CSPC response behaviours

7.2.1 Questionnaire completion rates

The number of individuals who were randomised to a DCE or CSPC questionnaire and the number who submitted a completed questionnaire are shown in Table 7.1, stratified by respondent subgroup.

Table 7.1: Completion rates by questionnaire format and respondent subgroup

| Survey sample | DCE | CSPC | p-value | Adjusted-p | Sig |
|-------------------|---------------|---------------|---------|------------|-----|
| All respondents | 656/738 (89%) | 662/792 (84%) | 0.003 | 0.009 | ** |
| General public | 595/640 (93%) | 595/672 (89%) | 0.007 | 0.014 | ** |
| Agent invitations | 61/98 (62%) | 67/120 (56%) | 0.410 | 0.410 | |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+'

The agent invitations sample includes individuals not self-identifying as decision-making agents. The adjusted p-values for the difference between the general public and agent invitation response rates were <0.001 in both the DCE and CSPC.

The completion rate among the general population subset was slightly but significantly higher in the DCE than the CSPC, while the difference among respondents to the agent invitations was not significantly different. Within each format, agents were significantly less likely than the general population sample to submit a completed questionnaire. Overall, the DCE had a significantly higher completion rate, although the absolute difference was only 5.3 percent.

As noted earlier, the invited sample of agents included individuals that were not necessarily decision-making agents, but 101 of the 128 total respondents to these invitations (79%) did self-identify as an agent, including 57 respondents to the CSPC (56%) and 44 respondents to the DCE (44%). Although an agent-specific denominator was not available, the relatively even distribution of agents across the two surveys contradicts the substantially lower representation of agents in the pilot CSPC, suggesting that agents were no more likely to drop out of the primary CSPC than the DCE.

The 'no answer' option was chosen in 5.8 percent of all DCE responses, and the rate was virtually identical among agents and the general public (5.1% vs. 5.8%, respectively, $p=0.37$). Linear regression found no significant trend in the proportion of no answers by task sequence ($p=0.98$), and contrary to the

hypothesis of greater utility balance in the repeated tasks leading respondents to opt out at a greater rate, the proportion of no answers in the repeated tasks was slightly but significantly less than the proportion across the other tasks combined (5.1% vs. 6.0%, $p=0.03$). If the greater utility balance in the repeated tasks did indeed result in a relatively more complex task, it does not appear that this complexity induced a greater proportion of no answers.

7.2.2 Respondent-rated difficulty and confidence

The proportions of respondents rating the tasks as somewhat or extremely difficult to understand are shown in Table 7.2, by format and respondent group. A greater proportion of respondents found the CSPC difficult to understand, particularly among agents, who were three times more likely to rate the CSPC as difficult compared to the DCE, although none of the differences were significant after adjusting for multiple comparisons. Within each format, the proportions of agents rating the tasks as difficult to understand were not significantly different from those of the general public.

Table 7.2: Respondents rating the tasks as somewhat or extremely difficult to understand, by format

| Group | DCE | CSPC | p-value | Adjusted-p | Sig |
|-----------------|---------------|---------------|---------|------------|-----|
| All respondents | 112/656 (17%) | 143/662 (22%) | 0.04 | 0.08 | + |
| General public | 108/612 (18%) | 126/605 (21%) | 0.18 | 0.18 | |
| Agents | 4/44 (9%) | 17/57 (30%) | 0.02 | 0.07 | + |

Significance codes: <0.001= '****' <0.01= '***' <0.05= '**' <0.10= '+'

The adjusted p-value for the difference in the proportions of the general public and agents rating the tasks as somewhat or extremely difficult was 0.85 in the DCE and 0.63 in the CSPC.

The proportions of respondents rating the tasks as somewhat or extremely difficult to answer are shown in Table 7.3. There were no significant differences in the proportions of agents or the general public who rated the tasks as difficult to answer, and no difference between agents and the general public in the proportion rating the DCE as difficult to answer, but a significantly greater proportion of agents rated the CSPC as difficult to answer compared to the public.

Table 7.3: Respondents rating the tasks as somewhat or extremely difficult to answer, by format

| Group | DCE | CSPC | p-value | Adjusted-p | Sig |
|-----------------|---------------|---------------|---------|------------|-----|
| All respondents | 311/656 (47%) | 328/662 (50%) | 0.47 | 0.81 | |
| General public | 284/612 (46%) | 286/605 (47%) | 0.81 | 0.81 | |
| Agents | 27/44 (61%) | 42/57 (74%) | 0.27 | 0.71 | |

Significance codes: <0.001= '****' <0.01= '***' <0.05= '**' <0.10='+'

The adjusted p-value for the difference in the proportions of the general public and agents rating the tasks as somewhat or extremely difficult was 0.31 in the DCE and 0.001 in the CSPC.

Finally, the proportions of respondents who indicated that they were somewhat or extremely confident that their answers in the DCE or CSPC choice tasks accurately reflected their preferences are shown in Table 7.4. A majority of all respondents were confident that their answers accurately represented their preferences, and there were no statistically significant differences between the two formats or between agents and the general public after adjusting for multiple comparisons.

Table 7.4: Respondents who indicated they were somewhat or extremely confident that their answers accurately reflected their preferences, by format

| Group | DCE | CSPC | p-value | Adjusted-p | Sig |
|-----------------|---------------|---------------|---------|------------|-----|
| All respondents | 461/656 (70%) | 431/662 (65%) | 0.05 | 0.21 | |
| General public | 429/612 (70%) | 393/605 (65%) | 0.06 | 0.26 | |
| Agents | 32/44 (73%) | 38/57 (67%) | 0.66 | 0.91 | |

Significance codes: <0.001= '****' <0.01= '***' <0.05= '**' <0.10='+'

The adjusted p-value for the difference in the proportions of the general public and agents rating their confidence as very or somewhat confident was 0.91 for both the DCE and the CSPC.

The relatively high confidence in both formats and in both respondent groups seemed to imply that respondents felt they were able to express their preferences accurately, regardless of any difficulties in understanding or answering the tasks. Indeed, a substantial proportion of the respondents who rated the tasks as difficult to answer also indicated that they were confident their answers accurately reflected their preferences.

7.2.3 Non-satiation and preference stability

Table 7.5 shows the proportion of respondents to the DCE and CSPC questionnaires who demonstrated a preference for the alternative with the dominant health outcomes in the test of non-satiation, by format and respondent subgroup. The results suggested that general population respondents to the DCE were substantially and significantly more likely to choose the dominant alternative compared to CSPC respondents (adjusted-p < 0.001), as almost all of the DCE respondents chose the dominant alternative compared to just over two-thirds of CSPC respondents. Eleven percent of all CSPC respondents equalised the budget allocation in this task and were not counted as prioritising the dominant alternative, but this only explains 41 percent of the relative difference between DCE and CSPC respondents. Agent respondents to the DCE survey also appeared substantially more likely to choose the dominant alternative compared to agents in the CSPC, although this difference was not as large as in the general population sample and was not statistically significant (adjusted-p=0.33). Agents were significantly less likely than general population respondents to choose the dominant alternative in the DCE (adjusted-p=0.003), but there was no significant difference between agents and the general public in the CSPC (adjusted-p=0.96).

Table 7.5: Non-satiation by questionnaire format and stakeholder group

| Stakeholder group | DCE | CSPC | p-value | Adjusted-p | Sig |
|-------------------|---------------|---------------|---------|------------|-----|
| All respondents | 625/656 (95%) | 457/662 (69%) | <0.001 | <0.001 | *** |
| General public | 588/612 (96%) | 417/605 (69%) | <0.001 | <0.001 | *** |
| Agents | 37/44 (84%) | 40/57 (70%) | 0.16 | 0.33 | |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

As noted earlier, stated preference elicitations generally assume that rational respondents should prefer the dominant alternative, but the normative quality of non-satiation in this context is not clear. Non-satiation ensures well-behaved, monotonically increasing indifference curves, but it is not a specific requirement of rationality in conventional choice theory (Lancsar & Louviere 2006). Indeed, a number of studies identified in the empirical ethics review found a preference for patients with the worst prospects, even if they would

benefit less from treatment than other groups. Therefore, although the difference in non-satiation between the two formats is notable, and may suggest that the CSPC induces a different cognitive process than the DCE, it is not in itself a fundamental advantage or limitation of either format.

Table 7.6 shows the proportion of respondents to the DCE and CSPC who were consistent in their preference for the same program in the original and the repeated task of each questionnaire. It suggests that respondents to the DCE had significantly greater preference stability, with a 10 percent advantage over the CSPC in all groups, although this difference was not statistically significant among the agent sample. The difference in the proportion of consistent responses between agents and the general public was not significant in the DCE ($p=0.41$) or the CSPC ($p=0.37$).

Table 7.6: Preference stability by questionnaire format and respondent subgroup

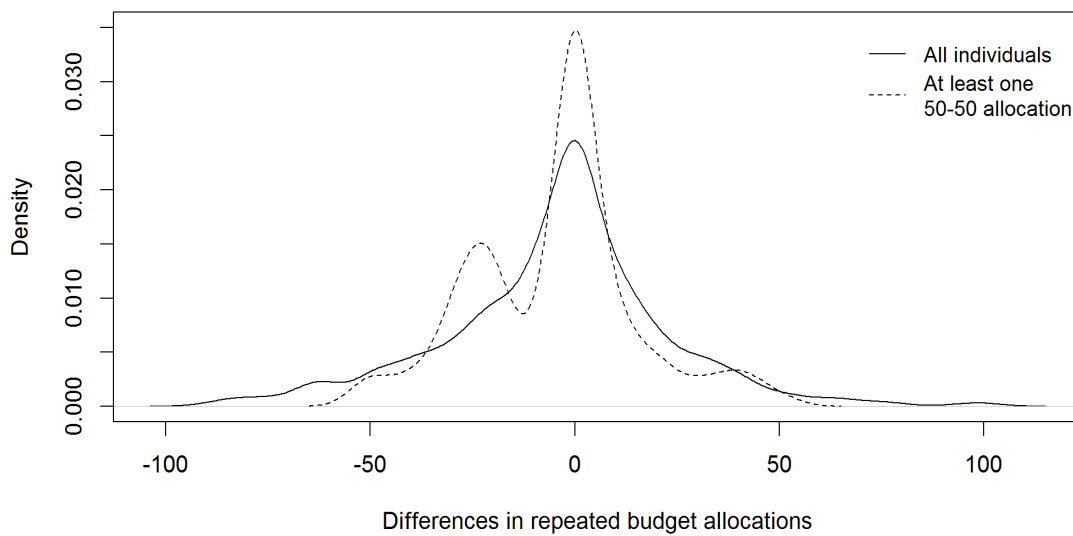
| Stakeholder group | DCE | CSPC | p-value | Adjusted p | Sig |
|-------------------|---------------|---------------|---------|------------|-----|
| All respondents | 475/656 (73%) | 414/662 (63%) | <0.001 | <0.001 | *** |
| General public | 446/612 (73%) | 382/605 (63%) | <0.001 | <0.001 | *** |
| Agents | 29/44 (66%) | 32/57 (56%) | 0.429 | 0.429 | |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

The adjusted-p for the difference between agents and the public in the proportion with stable preferences was 0.41 in the DCE and 0.37 in the CSPC.

The distribution of individual budget differences between the original and repeated CSPC tasks is shown in Figure 7.1. The mean budget allocation to program B was 41.9 percent in the initial task and 37.7 percent to the same program in the repeated task. The individual differences were clustered around zero, confirming that most CSPC respondents allocated a roughly similar budget share in both tasks, and although the mean difference was significantly different from zero, it was still relatively small (mean absolute difference=-4.2%, $p<0.001$).

Figure 7.1: Distribution of individual budget differences between the repeated CSPC tasks

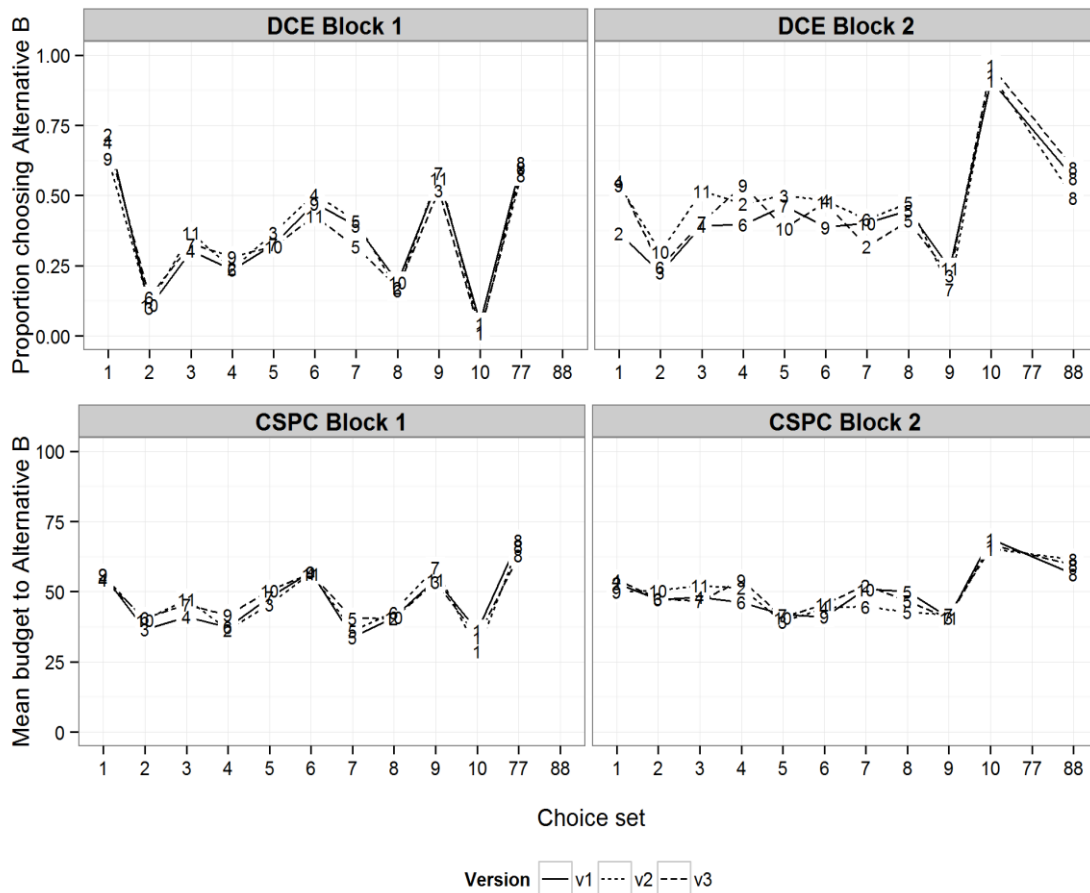


Ninety-six of the 662 CSPC respondents (15%) chose an equal 50-50 budget allocation in at least one of the original or repeated task, and 27 of those respondents were consistent in choosing an equal allocation in both tasks. As with the overall distribution, the budget differences for individuals with at least one 50-50 budget allocation were clustered around zero, and although the mean difference was statistically significant it was again relatively small (mean absolute difference=-4.8, $p=0.02$).

7.2.4 Learning and fatigue effects

The mean proportions of DCE respondents preferring alternative B in each choice set, and the mean CSPC budget allocations to alternative B in each choice set, stratified by format, block and questionnaire version, are shown in Figure 7.2. Note that the choice sets differed between blocks, and that the order that the choice sets were presented within each block differed by version. The numeric labels indicate the sequence in which the choice sets were presented in each version.

Figure 7.2: Choices and budget allocations by design, choice set and task sequence



The order the choice sets were presented differed by version, and the numeric labels indicate the order each choice set appeared in the different versions. Choice set 10 was the test of dominance and was always presented as the first task. Choice sets 77 and 88 were the reversed versions of sets 7 and 8, respectively. The original and repeated choice sets were always presented as tasks 5 and 8, respectively.

The figures suggest that there was very little difference in the proportions or means over the choice sets by version, and this impression was largely confirmed by the results of the one-way ANOVAs by format, block and choice set shown in Table 7.7.

Table 7.7: ANOVA adjusted p-values by choice set, format and block

| Choice set | DCE, Block 1 | DCE, Block 2 | CSPC, Block 1 | CSPC, Block 2 |
|------------|--------------|--------------|---------------|---------------|
| 1 | 0.393 | 0.008 | 0.812 | 0.386 |
| 2 | 0.706 | 0.405 | 0.428 | 0.517 |
| 3 | 0.614 | 0.113 | 0.273 | 0.269 |
| 4 | 0.694 | 0.114 | 0.269 | 0.063 |
| 5 | 0.705 | 0.200 | 0.441 | 0.781 |

| | | | | |
|----|-------|-------|-------|--------------|
| 6 | 0.510 | 0.274 | 0.987 | 0.396 |
| 7 | 0.325 | 0.273 | 0.132 | 0.058 |
| 8 | 0.866 | 0.630 | 0.825 | 0.090 |
| 9 | 0.626 | 0.365 | 0.284 | 0.921 |
| 10 | 0.212 | 0.297 | 0.104 | 0.545 |
| 77 | 0.801 | N/P | 0.228 | N/P |
| 88 | N/P | 0.259 | N/P | 0.251 |

P-values less than 0.10, highlighted in **bold**, indicate a significant difference in the proportions or means by questionnaire version. The order of the choice sets varied by version, but choice set 10 was the test of dominance and was always presented as the first task. Choice set 7 was repeated as set 77 in block 1 of both formats, and set 8 was repeated as set 88 in block 2 of both formats. The original and repeated tasks were always presented as tasks 5 and 8, respectively. N/P=Not presented in the design block.

Choice set 1 of DCE block 2 was the only set with a statistically significant difference at a 0.05 threshold, but relaxing this threshold to 0.10 added three other instances, all from block 2 of the CSPC: choice sets 4, 7 and 8. Post hoc tests of these four instances are shown in Appendix 7.1. Set 1 of DCE block 2, version 1, which presented this set as the second task in the sequence, was associated with a significantly lower probability of choice relative to versions 2 (task 9) and 3 (task 4), which presented the task later in the sequence. In CSPC block 2, set 4 had a higher mean budget allocation as task 9 than as task 6, while set 7 had a significantly higher mean budget allocation as task 2 than task 6. In both cases, earlier tasks had significantly greater budget allocations. There was also a statistically significant difference in set 8 of CSPC block 2, but recall that was the original task in the repeated test of preference consistency and was presented at the same point in the task sequence (task 5) in all three versions. This suggests that this observed difference was not specifically associated with learning or fatigue effects, although it is possible that there were more complex ordering effects in that block.

These differences suggest the possibility of learning effects in both the DCE and the CSPC, as responses to the tasks presented earlier in the choice sequence were significantly different from responses when the same sets were presented later in the sequence. However, only 4 out of the 44 possible choice sets were associated with any statistically significant differences, and in the case of the CSPC these differences were relatively small (absolute differences of 5-8%). It is difficult to conclude, therefore, that there were meaningful learning or fatigue effects over the sequence of choice tasks presented in either format. This

seems particularly noteworthy for the more complex and cognitively demanding CSPC.

7.2.5 Dominant preferences

The distribution of individual choice-attribute correlations by format and attribute are illustrated in Appendix 7.3. In these histograms, a perfect correlation between choice and the *higher* level of an attribute appears as a correlation coefficient of 1.0, and a perfect correlation between choice and the *lower* level of an attribute appears as -1.0. The proportion of respondents with at least one perfect correlation between choice and a particular attribute flag is shown in Table 7.8, along with the proportions of respondents with perfect choice correlations by specific attributes. By chance, it was possible for respondents to have perfect choice correlations with more than one attribute, so the sum of respondents across attributes is greater than the number of unique individuals with a perfect-choice attribute correlation. As well, the alternative with the greatest number of patients treated was also always the alternative with the greatest aggregate QALYs gained. As the two attribute flags were themselves perfectly correlated, it was not possible to disentangle the choice correlations for the two attributes, and the same individuals were counted in both attributes.

Table 7.8: Individuals with perfect choice-attribute correlation by attribute and format

| Attribute | DCE | CSPC | p value | Adjusted-p | Sig |
|---|---------------|---------------|---------|------------|-----|
| Any perfect correlation | 61/656 (9.3%) | 24/658 (3.6%) | <0.001 | <0.001 | *** |
| <i>By attribute</i> | | | | | |
| Age | 39/656 (5.9%) | 16/658 (2.4%) | 0.002 | 0.011 | * |
| Life expectancy | 3/656 (0.5%) | 1/658 (0.2%) | 0.610 | 1.000 | |
| Life years gained | 0/656 (0.0%) | 1/658 (0.2%) | 1.000 | 1.000 | |
| Patients treated | 19/656 (2.9%) | 4/658 (0.6%) | 0.003 | 0.015 | * |
| Aggregate QALYs | 19/656 (2.9%) | 4/658 (0.6%) | 0.003 | 0.015 | * |
| Initial health state | 0/656 (0.0%) | 2/658 (0.3%) | 0.483 | 1.000 | |
| Final health state | 0/656 (0.0%) | 0/658 (0.0%) | n/d | n/d | |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+' | | | | | |

n/d: p-value not defined where both formats have no events. The CSPC results exclude 4 respondents who did not move the slider in any of their choices and finished the questionnaire in less than one-half the median completion time.

The proportion of respondents with at least one perfect choice-attribute correlation in the DCE was significantly higher than the proportion in the CSPC (9.3% vs. 3.6%, adjusted-p <0.001). Consistent with this overall difference, there were also significant differences between formats in the proportions with a perfect choice correlation with age, total patients treated and aggregate QALYs gained. The majority of perfect choice-attribute correlations were associated with the age attribute, where 12 out of 16 CSPC respondents (67%) and 33 out of 39 DCE respondents (85%) favoured the lower (younger) level. All 19 DCE respondents who had a perfect choice correlation with total patients and aggregate QALYs favoured higher levels, but 2 of the 4 CSPC respondents who had a perfect correlation with these attributes always chose the lower levels.

To distinguish dominant preferences from perfect choice-attribute correlations that may have happened by chance, each perfect correlation was compared to the respondent's self-rated importance ratings. Table 7.9 shows the proportion of respondents with at least one perfect choice-attribute correlation who also ranked that attribute as most important in the rating task that followed the DCE and CSPC choice tasks. As it was possible for a respondent to have more than one perfect choice-attribute correlation, and to rate multiple attributes as equally important, it was possible for a respondent to have a confirmed dominant preference for more than one attribute. This was particularly true for

the number of patients treated and aggregate QALYs gained, where as noted above, the flags for the two attributes were themselves perfectly correlated.

Table 7.9: Individuals with confirmed dominant preferences by attribute and format

| Attribute | DCE | CSPC | p-value | Adjusted-p | Sig |
|--|---------------|---------------|---------|------------|-----|
| Any dominant preference | 45/656 (6.9%) | 18/658 (2.7%) | <0.001 | 0.006 | ** |
| <i>By attribute</i> | | | | | |
| Age | 32/656 (4.9%) | 14/658 (2.1%) | 0.010 | 0.049 | * |
| Life expectancy | 0/656 (0.0%) | 0/658 (0.0%) | n/d | n/d | |
| Life years gained | 0/656 (0.0%) | 1/658 (0.2%) | 1.000 | 1.000 | |
| Patients treated | 13/656 (2.0%) | 2/658 (0.4%) | 0.009 | 0.045 | * |
| Aggregate QALYs | 15/656 (2.3%) | 2/658 (0.4%) | 0.003 | 0.022 | * |
| Initial health state | 0/656 (0.0%) | 0/658 (0.0%) | n/d | n/d | |
| Final health state | 0/656 (0.0%) | 0/658 (0.0%) | n/d | n/d | |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+' | | | | | |

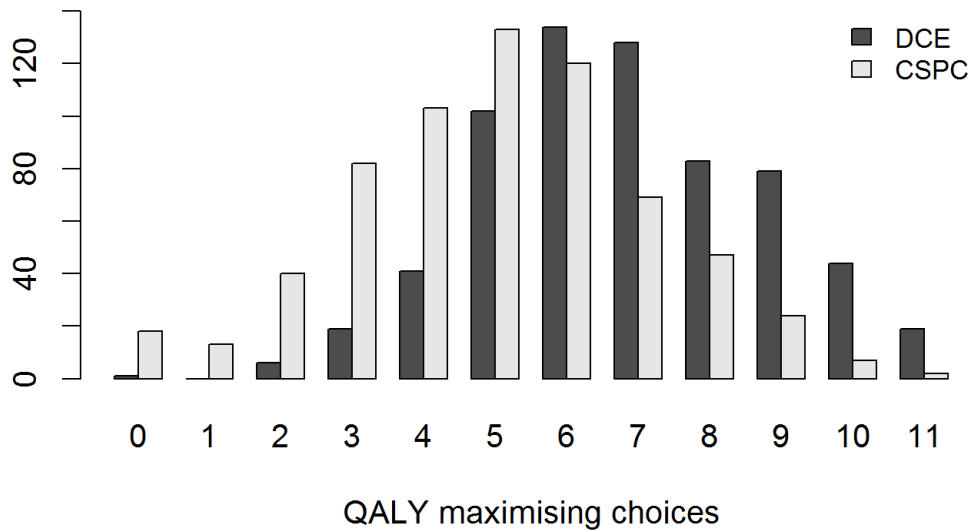
n/d: p-value not defined where both formats have no events. The CSPC results exclude 4 respondents who did not move the slider in any of their choices and finished the questionnaire in less than one-half the median completion time.

Again, the DCE was associated with a significantly higher proportion of respondents with a confirmed dominant preference (6.9% vs. 2.7%, adjusted-p=0.006). There were also significant differences between the formats in the proportion of respondents holding a dominant preference for age, total patients treated and aggregate QALYs gained, with the DCE higher across all three attributes. The most common dominant preference in both formats was for age, where 10 out of the 14 CSPC respondents (71%) and 30 out of the 32 DCE respondents (94%) had a dominant preference for the younger patient group in each choice. This was in contrast to the pilot survey, where the most frequent dominant preference in the DCE was for final health state, and in the CSPC was for individual life years gained.

7.2.6 QALY maximisation

The distribution of respondents by the count of their total QALY maximising choices out of the 11 choice tasks in each questionnaire is illustrated in Figure 7.3, and detailed in Table 7.10.

Figure 7.3: QALY maximising choices by questionnaire format



Contrary to the hypothesis of a possible prominence effect in the CSPC, which may have encouraged respondents to give more weight to aggregate QALY gains in their allocations, it was DCE respondents who were significantly more likely to choose the QALY maximising alternatives in their choice tasks. DCE respondents made on average almost two more QALY maximising choices than CSPC respondents ($p < 0.001$). Overall, there was little evidence of QALY maximising behaviour among respondents to either format, as only 2 percent of all respondents prioritised the alternative that maximised QALYs in every task. It is interesting to note that a majority of CSPC respondents (59%) consistently prioritised the QALY *minimising* alternative in making five or fewer QALY maximising choices, and this was significantly more than the proportion of DCE respondents (26%, $p < 0.001$).

Table 7.10: Respondents by number of QALY maximising choices and questionnaire format

| QALY maximising choices | DCE | CSPC | Combined |
|-------------------------|---------|----------|----------|
| 0 | 1 (0%) | 18 (3%) | 19 (1%) |
| 1 | 0 (0%) | 13 (2%) | 13 (1%) |
| 2 | 6 (1%) | 40 (6%) | 46 (3%) |
| 3 | 19 (3%) | 82 (12%) | 101 (8%) |

| | | | |
|--|------------------|------------------|--------------------|
| 4 | 41 (6%) | 103 (16%) | 144 (11%) |
| 5 | 102 (16%) | 133 (20%) | 235 (18%) |
| 6 | 134 (20%) | 120 (18%) | 254 (19%) |
| 7 | 128 (20%) | 69 (10%) | 197 (15%) |
| 8 | 83 (13%) | 47 (7%) | 130 (10%) |
| 9 | 79 (12%) | 24 (4%) | 103 (8%) |
| 10 | 44 (7%) | 7 (1%) | 51 (4%) |
| 11 | 19 (3%) | 2 (0%) | 21 (2%) |
| All respondents | 656 (50%) | 658 (50%) | 1314 (100%) |
| Mean QALY maximising choices per respondent | 6.81 | 5.02 | 5.91 |
| p-value | <0.001 | | |

The CSPC results exclude 4 respondents who did not move the slider in any of their choices and finished the questionnaire in less than one-half the median completion time.

Table 7.11 shows the distribution of respondents by their QALY maximising choices, this time stratified by agent status. The number of QALY maximising choices made by agents was not significantly different than the number made by respondents from the general population sample (5.83 vs. 5.92, $p=0.70$).

Table 7.11: Respondents by number of QALY maximising choices and agent status

| QALY maximising choices | Agents | Public | Combined |
|--|-----------------|-------------------|--------------------|
| 0 | 0 (0%) | 19 (2%) | 19 (1%) |
| 1 | 2 (2%) | 11 (1%) | 13 (1%) |
| 2 | 6 (6%) | 40 (3%) | 46 (3%) |
| 3 | 8 (8%) | 93 (8%) | 101 (8%) |
| 4 | 11 (11%) | 133 (11%) | 144 (11%) |
| 5 | 13 (13%) | 222 (18%) | 235 (18%) |
| 6 | 26 (26%) | 228 (19%) | 254 (19%) |
| 7 | 11 (11%) | 186 (15%) | 197 (15%) |
| 8 | 11 (11%) | 119 (10%) | 130 (10%) |
| 9 | 10 (10%) | 93 (8%) | 103 (8%) |
| 10 | 2 (2%) | 49 (4%) | 51 (4%) |
| 11 | 1 (1%) | 20 (2%) | 21 (2%) |
| All respondents | 101 (8%) | 1213 (92%) | 1314 (100%) |
| Mean QALY maximising choices per respondent | 5.83 | 5.92 | 5.91 |
| p-value | 0.70 | | |

The public results exclude 4 CSPC respondents who did not move the slider in any of their choices and finished the questionnaire in less than one-half the median completion time.

Finally, as the individual-level QALY graphs presented to respondents may have encouraged a focus on individual rather than aggregate QALY gains, Table 7.12 shows the distribution of respondents by the number of choices they made that maximised individual QALY gains and aggregate QALYs gains. The table shows that respondents were slightly but significantly more likely to choose the alternative that maximised aggregate QALY gains over individual QALY gains (5.91 vs. 5.72, $p=0.02$). This suggests that the individual-level QALY graphs did not focus respondent attention on individual gains to the exclusion of consideration of aggregate gains. Similar to the result observed in Table 7.10, when results shown in Table 7.12 were further stratified by questionnaire format (not shown), the majority of CSPC respondents (56%) were more likely to prioritise the individual QALY *minimising* alternative, compared to 36 percent of DCE respondents ($p<0.001$). In considering these results, note that the individual QALY maximising and the aggregate QALY maximising alternatives were often one and the same. The results shown in Table 7.12 should therefore be interpreted in terms of the relative trend rather than as an absolute trade-off between individual or aggregate QALY gains.

Table 7.12: Respondents by number of individual and aggregate QALY maximising choices

| QALY maximising choices | Individual QALYs | Aggregate QALYs |
|--|--------------------|--------------------|
| 0 | 13 (1%) | 19 (1%) |
| 1 | 11 (1%) | 13 (1%) |
| 2 | 36 (3%) | 46 (4%) |
| 3 | 108 (8%) | 101 (8%) |
| 4 | 187 (14%) | 144 (11%) |
| 5 | 251 (19%) | 235 (18%) |
| 6 | 246 (19%) | 254 (19%) |
| 7 | 220 (17%) | 197 (15%) |
| 8 | 127 (10%) | 130 (10%) |
| 9 | 80 (6%) | 103 (8%) |
| 10 | 27 (2%) | 51 (4%) |
| 11 | 8 (1%) | 21 (2%) |
| All respondents | 1314 (100%) | 1314 (100%) |
| Mean QALY maximising choices per respondent | 5.72 | 5.91 |
| p-value | 0.021 | |

Excludes 4 CSPC respondents who did not move the slider in any of their choices and finished the questionnaire in less than one-half the median completion time.

Overall, respondents chose the aggregate QALY maximising alternative in just over half of all tasks, and 58 percent of all respondents maximised QALYs in more than half of their choices. However, the Monte Carlo simulated confidence intervals suggested that the difference between the number of QALY maximising choices made by agents was not significantly different than the number that might be expected by chance (mean difference=0.11; 95% CI: -0.21, 0.44). Simulated confidence intervals for the general population sample were positive and statistically significantly different than chance (mean difference=0.29; 95% CI: 0.20, 0.37), which suggested some support for QALY maximisation, although the size and meaningfulness of this difference was still quite small.

A probit model tested the impact of task sequence, questionnaire format, and agent status, as well as interactions between sequence and format, and sequence and agent status on the likelihood of choosing the QALY maximising alternative in each task. After adjusting for clustering in the standard errors, agent status and the two interaction terms were not statistically significant at a 0.10 threshold, and the model was re-estimated with task sequence and questionnaire format only. The results of this more parsimonious model, shown in Appendix 7.2, suggested a statistically significant negative trend in the likelihood of choosing the QALY maximising alternative as a respondent progressed through the task sequence. As well, consistent with the significant difference in mean number of QALY maximising choices by format shown above, the CSPC format was associated with a significantly lower likelihood of choosing the QALY maximising alternative. The predicted choice probabilities over the task sequence by questionnaire format are shown in Table 7.13.

Table 7.13: Predicted QALY maximising probabilities by task sequence and format

| Task sequence | DCE | 95% CI | CSPC | 95% CI |
|---------------|-------|-----------------|-------|-----------------|
| 2 | 0.618 | (0.600 - 0.635) | 0.466 | (0.448 - 0.484) |
| 3 | 0.611 | (0.595 - 0.626) | 0.459 | (0.443 - 0.475) |
| 4 | 0.604 | (0.590 - 0.617) | 0.452 | (0.438 - 0.466) |
| 5 | 0.597 | (0.584 - 0.609) | 0.444 | (0.432 - 0.457) |
| 6 | 0.589 | (0.577 - 0.601) | 0.437 | (0.425 - 0.449) |
| 7 | 0.582 | (0.570 - 0.594) | 0.430 | (0.418 - 0.442) |
| 8 | 0.575 | (0.562 - 0.588) | 0.423 | (0.410 - 0.436) |

| | | | | |
|----|-------|-----------------|-------|-----------------|
| 9 | 0.568 | (0.554 - 0.582) | 0.416 | (0.402 - 0.430) |
| 10 | 0.561 | (0.545 - 0.576) | 0.408 | (0.393 - 0.424) |

Task 1 was always the test of dominance, and was excluded from the model as a possible outlier in the overall trend. 95% CI = 95% confidence interval.

These probabilities illustrated a slight but statistically significant downward trend in the likelihood of choosing the QALY maximising alternative over the sequence of tasks, independent of the attribute levels in the tasks themselves, which varied by questionnaire block and version. There was no overlap in the confidence intervals between the formats, suggesting a significantly lower likelihood of CSPC respondents prioritising the QALY maximising alternative in any given task.

These results appeared consistent with the hypothesis that respondents become less likely to prioritise on the basis of aggregate QALY gains as they became more familiar with the trade-offs and attribute levels in the choice tasks. They may also be consistent with a similar idea that the initial test of non-satiation primed respondents to favour the QALY maximising alternative, but that this effect gradually wore off over the course of the choice tasks. Overall, the relatively small effect over the range of tasks tested suggested that the net impact of this trend on preferences would be minimal.

7.3 Discussion of the DCE-CSPC comparisons

The results of the DCE-CSPC comparisons are summarised in Table 7.14, and in general they reinforce the findings of the pilot study. The DCE appeared to be associated with greater respondent efficiency in terms of better completion rates and greater preference consistency, while the CSPC appeared to be a more cognitively demanding task, associated with longer completion times and more difficulty understanding the task, particularly among agents.

Table 7.14: Summary of DCE-CSPC comparisons

| Comparison | DCE | CSPC | Adjusted-p | Sig |
|---|-------------|--------------|------------|-----|
| Overall completion rate | 89% | 84% | 0.009 | ** |
| Median completion time | 9.5 minutes | 11.7 minutes | <0.001 | *** |
| Fast completers | 9.1% | 11.3% | 0.22 | |
| Somewhat or extremely difficult to understand | 17% | 22% | 0.08 | + |

| | | | | |
|--|------|------|--------|-----|
| Somewhat or extremely difficult to answer | 47% | 50% | 0.81 | |
| Somewhat or extremely confident choices reflect preferences | 70% | 65% | 0.21 | |
| Non-satiation (% preferring dominant alternative) | 95% | 69% | <0.001 | *** |
| Stable preferences in repeated task | 73% | 63% | <0.001 | *** |
| Perfect choice-attribute correlations | 9.3% | 3.6% | <0.001 | *** |
| Confirmed dominant preferences | 6.9% | 2.7% | 0.006 | ** |
| Mean QALY-maximising choices (out of 11) | 6.81 | 5.02 | <0.001 | *** |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+' | | | | |

Some agents commented that the CSPC task, which asked respondents to divide a fixed budget between two alternatives, was not a realistic reflection of their usual decision-making tasks. This highlights the fact that respondents are used to making decisions within a particular context, and as noted in Chapter 4, changing the context of a decision to suit a particular elicitation method may adversely impact the face validity, accuracy and predictive ability of a task.

The longer median completion time and the lower incidence of perfect choice-attribute correlations and dominant preferences among CSPC respondents, including for the QALY maximising alternative, are also consistent the characterisation of the CSPC as a more reflective task than the DCE. The CSPC required respondents to consider the relative value of the two alternatives in each task, and this may encourage them to consider the overall quality of both alternatives to a greater degree than the 'pick one' nature of the DCE (Carson & Louviere 2011; Huber 2009). Schwappach and Strasmann (2006) argue that the ability to reserve a portion of the budget for a less preferred group will also tend to make CSPC tasks more reflective, as respondents must consider how much of the budget, if any, to reserve. This explicit consideration of the less preferred group may also explain the greater proportion of CSPC respondents (31%) who gave priority to the non-dominant group in the test of non-satiation compared to the DCE (5%). Some of this difference is explained by the 11 percent of CSPC respondents who chose to equalise the budget allocations rather than prioritise one group or the other, but the remainder appeared to reflect a fundamentally different cognitive process in the CSPC compared to the DCE, perhaps leading to a relatively greater concern for patients with poorer prospects and/or potential

for gain. Qualitative work to understand the specific rationale of CSPC and DCE respondents in this task would be of interest, particularly in confirming what might be termed a 'compassion bias' in the CSPC.

There was little evidence of strict QALY maximising behaviour among respondents to either format, as only 2 percent of all respondents prioritised the alternative that maximised QALYs in every task. Respondents to the CSPC were significantly less likely than respondents to the DCE to choose the QALY maximising alternative, consistent with the notion of a compassion bias in the CSPC. Again, though, some of the lower rate of QALY maximising behaviour may be explained by the opportunity CSPC respondents had to equalise allocations between alternatives. Agents were also no more likely than the general population to be strict QALY maximisers. In fact, the number of QALY maximising choices made by agents was slightly but significantly less than the general population sample, and was not significantly different than what would be expected by chance. Some of the higher rate of QALY maximising behaviour among the general population sample may have been the result of a simplifying QALY-maximising decision rule, while the lower rate observed among agents may have been driven by the relatively high proportion of clinicians among the agent sample: almost two-thirds of the agents in the sample were oncology professionals and they may have been less likely than non-clinicians to adopt a QALY maximising rule.

Although the CSPC makes the trade-offs between patient groups more explicit, as respondents see the number of patients treated in one group decline as they allocate resources to the other, this did not appear to translate into a significant prominence effect around this attribute or aggregate QALYs gained that might encourage respondents to maintain a societal-level perspective. Furthermore, the higher completion rate and similar levels of preference confidence in the DCE relative to the CSPC do not appear to support Swallow et al.'s (2001) contention that respondents may be reluctant to complete dichotomous preference elicitation over highly emotive choices. In the absence of these hypothesised advantages of CSPC, the greater completion rate and slightly more favourable difficulty rating of the DCE gives it a pragmatic advantage for the elicitation of societal preferences. However, the CSPC was

associated with a significantly lower incidence of dominant preferences, and the cardinal nature of its response format suggests that it may have an advantage in terms of statistical efficiency, although this was not specifically tested here. Overall, both formats were associated with similar difficulty ratings and preference confidence, minimal learning or fatigue effects, and relatively few cases of dominant preferences. This suggests that both formats are eliciting valid preference data, which will be presented over the next two chapters.

Appendix 7.1: Post hoc test of significant ANOVA results

Table 7.15: Tukey's test of honest significant difference

| Version (Task sequence) | Difference | L95CI | U95CI | Adjusted-p | Sig |
|--|------------|---------|--------|------------|-----|
| DCE block 2, set 1 | | | | | |
| v2(9) - v1(2) | 0.172 | 0.018 | 0.325 | 0.024 | * |
| v3(4) - v1(2) | 0.185 | 0.028 | 0.343 | 0.016 | * |
| v3(4) - v2(9) | 0.014 | -0.142 | 0.169 | 0.977 | |
| CSPC block 2, set 4 | | | | | |
| v2(2) - v1(6) | 5.087 | -3.270 | 13.443 | 0.325 | |
| v3(9) - v1(6) | 7.904 | -0.030 | 15.837 | 0.051 | + |
| v3(9) - v2(2) | 2.817 | -5.200 | 10.832 | 0.686 | |
| CSPC block 2, set 7 | | | | | |
| v2(6) - v1(10) | -6.140 | -14.253 | 1.972 | 0.177 | |
| v3(2) - v1(10) | 1.481 | -6.220 | 9.183 | 0.893 | |
| v3(2) - v2(6) | 7.622 | -0.160 | 15.404 | 0.056 | + |
| CSPC block 2, set 8 | | | | | |
| v2(5) - v1(5) | -7.306 | -15.126 | 0.514 | 0.073 | + |
| v3(5) - v1(5) | -3.426 | -10.850 | 3.998 | 0.523 | |
| v3(5) - v2(5) | 3.880 | -3.621 | 11.381 | 0.444 | |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+' | | | | | |

Table shows the pairwise comparison of differences in the proportion of DCE respondents preferring alternative B in block 2, set 1 by questionnaire version. L95CI=Lower 95% confidence interval; U95CI=Upper 95% confidence interval

Appendix 7.2: Probit analysis of QALY maximising choices by task sequence and questionnaire format

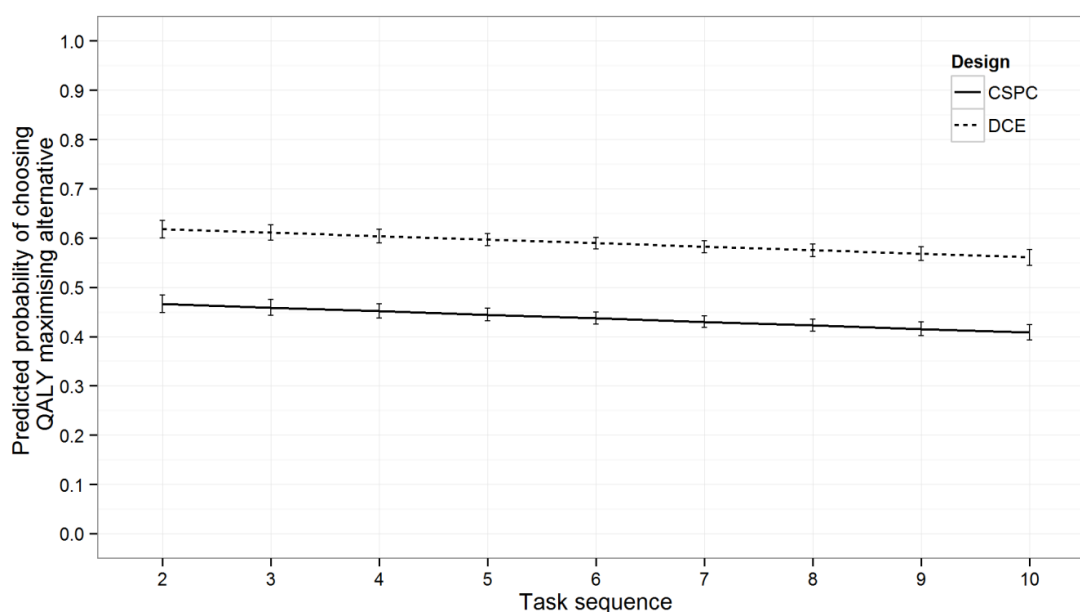
Table 7.16: Probit model of likelihood of choosing the QALY maximising alternative

| Factor | Estimate | Std. Error* | z value | Pr(> z) | Sig |
|--------------|----------|-------------|---------|-----------|-----|
| (Intercept) | 0.336 | 0.029 | 11.460 | <0.001 | *** |
| Sequence | -0.018 | 0.004 | -4.811 | <0.001 | *** |
| Format: CSPC | -0.384 | 0.022 | -17.457 | <0.001 | *** |

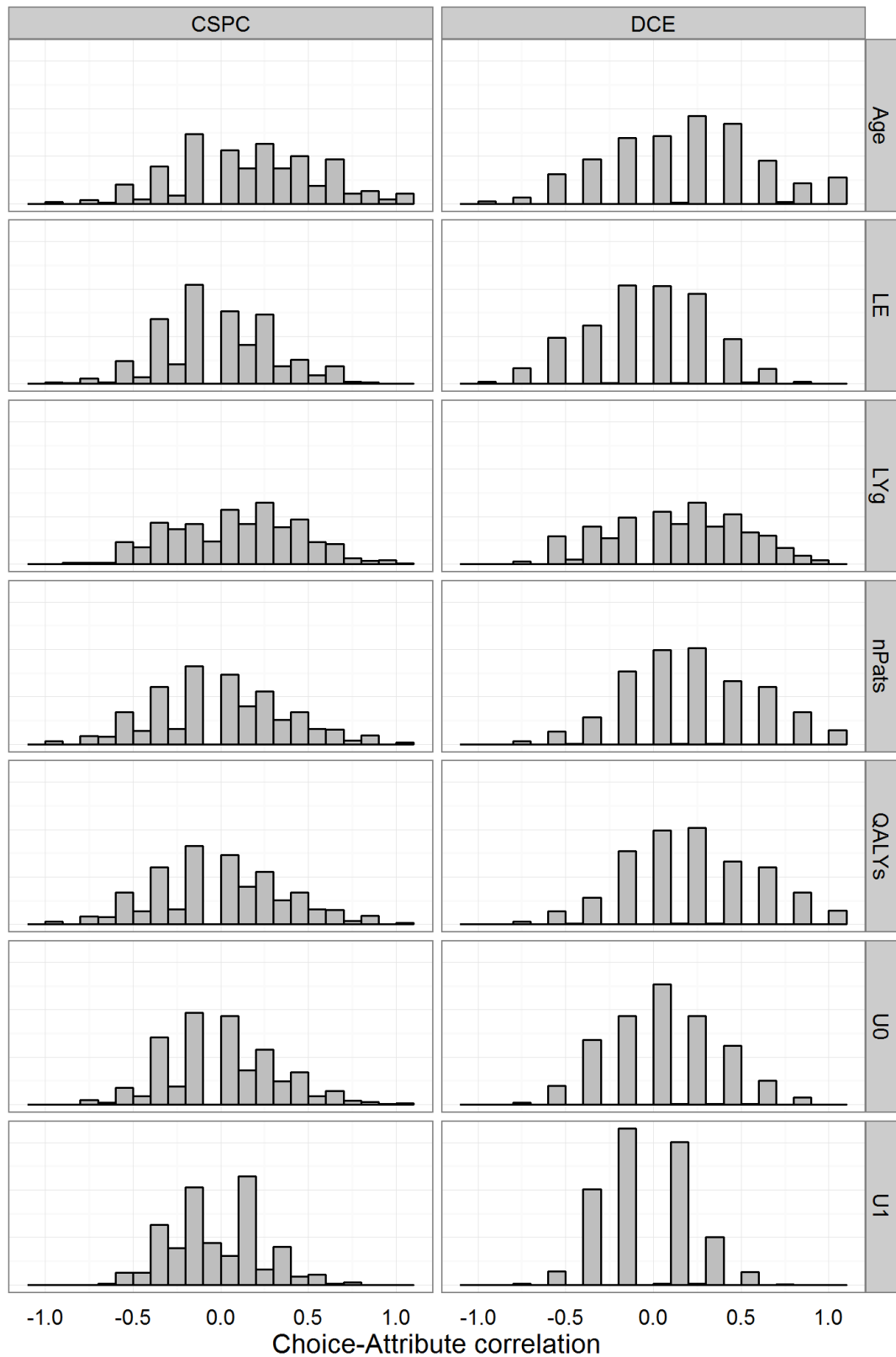
Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

* Standard errors were adjusted for clustering using the 'sandwich estimator'. (Freedman 2006; Zeileis & Hothorn 2002)

Figure 7.4: Predicted probabilities of choosing/prioritising the QALY maximising alternative by task and questionnaire format



Appendix 7.3: Distribution of choice-attribute correlations



Age=patient age; LE=initial life expectancy; LYg=individual life years gained; nPats=number of patients treated; QALYs=aggregate QALYs gained; U0=initial utility; U1=final utility. Correlations towards -1.0 indicate a lower level in an attribute was consistently preferred, and correlations towards 1.0 indicate a higher level in an attribute was consistently preferred.

Chapter 8: Primary DCE results

The primary objective of the analysis of the DCE responses was to estimate the strength of the equity-efficiency trade-off between individual life year gains and the equity factors identified in the empirical ethics review: age, initial health state, untreated life expectancy, and final health state, as well as the relative distribution of benefits. To use Broome's (1989) terminology, these factors were taken to constitute *claims* to scarce healthcare resources on the basis of empirical support and defensible ethical justifications, distinct from the wider set of *reasons* that a particular patient or group might deserve priority. The relative strength of the trade-off over different attributes was taken as an estimate of the strength of societal preferences, or the welfare effect, associated with prioritising patients or groups with those particular characteristics. From a Communitarian perspective, prioritising patient groups that best satisfy community preferences is argued to increase overall societal well-being.

Section 8.1 describes the specification of the DCE choice model, including the rationale for choosing between additive versus multiplicative and linear versus effects-coded value functions. Section 8.2 outlines the issue of individual heterogeneity in preferences and describes the two leading methods for dealing with this issue in the context of a DCE: latent class models and random parameters models. The use of compensating variation as an estimate of the strength of the equity-efficiency trade-off is discussed in section 8.3, including its advantage over the marginal rate of substitution, which was used in the pilot survey. The results of the DCE model, and the derived welfare effects, are described in section 8.4, along with a ranking of choice scenarios by predicted

utility and a comparison of agent and public preferences. Finally, these results are interpreted and discussed in section 8.5.

8.1 Specifying the DCE model

The analysis of the DCE responses was based on the assumption that respondents derived different degrees of utility from allocating resources to patient groups with different characteristics or attributes. An additive main effects value function was specified for the experimental design, but alternative value functions suggested by the literature were also considered, including main effects interacted with life year gains and a multiplicative log-linear specification.

In its simplest form, an additive value function implies that the utility derived from each attribute is independent of the level of the other attributes. Any utility derived from allocating healthcare resources to a group of younger patients, for example, would be independent of other attributes such as life years gained with treatment or initial health state. This functional form is common in the DCE literature, and is consistent with recent models of choice in a societal healthcare context (Bryan et al. 2002; Mortimer & Segal 2008; Green & Gerard 2009; Ratcliffe et al. 2009). The experimental design was based on an additive value function of the form $v = \beta_1LYg + \beta_2Age + \beta_3U0 + \beta_4LE + \beta_5UI + \beta_6nPats + \beta_7UI \cdot LYg$, where LYg is the number life years gained per patient with treatment, Age is the average age of patients in the group, $U0$ is the quality of the initial health state, measured on a 0-1 utility scale, $LE0$ is life expectancy without treatment, UI is the utility of the health state with/after treatment, $nPats$ is the total number patients that could be treated, and $UI \cdot LYg$ is the interaction of UI and LYg , intended to account for the quality of additional life years. Aggregate QALYs gained, as a linear combination of the other attributes, was excluded to avoid collinearity. In addition, the use of the QALY pre-supposes a specific trade-off between life years gained and health state that may not hold in this context. The age and the number of patients treated parameters were divided by 10 and 1000, respectively, to re-scale them to a magnitude comparable with the other parameters in order to improve the chances of model convergence (Long 1997). An interaction term capturing the absolute change in utility, as $(1-U0)UI$,

was also incorporated in more complex versions of the additive value function. The possible values of this interaction term are shown below:

Table 8.1: Initial and final health state interaction values

| | | U1 | | | |
|-----|------|------|------|------|--|
| | | 0.1 | 0.5 | 0.9 | |
| U0 | 1-U0 | | | | |
| 0.1 | 0.9 | 0.09 | 0.45 | 0.81 | |
| 0.5 | 0.5 | 0.05 | 0.25 | 0.45 | |
| 0.9 | 0.1 | 0.01 | 0.05 | 0.09 | |

U0 = initial utility; U1 = final utility

The value of this term was maximised in scenarios when patients move from the worst initial health state to the best final health state, and minimised when patients move from the best initial health state to the worst final health state.

Norman et al. (2013) suggested that a strictly additive value function is inappropriate in the context of health programs, arguing that as the health gains derived from a hypothetical program tend to zero, so too should the utility associated with that program, regardless of the level of other attributes. This is analogous to the ‘zero condition’ of the QALY model, which implies that different health states with a duration of zero life years will all have zero utility, regardless of their quality or other characteristics (Miyamoto et al. 1998). As such, they used an additive value function but interacted each attribute with the gain in life expectancy in their analysis of relative preferences for efficiency and equity in the allocation of healthcare resources. In this form, the utility associated with different patient attributes is dependent on gains in life expectancy, with the other attributes weighting, positively or negatively, the net value of that gain. Lancsar et al. (2011) also made the argument that utility in the context of healthcare resource allocation should be dependent on health gain, but used a log-linear value function to model the utility as a multiplicative function of the logged attribute levels¹³ in order to estimate QALY distributional preferences. In this form, utility can be non-linear but still monotonic over the range of an attribute.

¹³ Recall that $\log(a) + \log(b) = \log(ab)$

However, the DCE elicitations in these two studies were structured differently than the elicitation presented here. Norman et al. (2013) elicited preferences over changes in life expectancy to patients described in terms of gender, smoking status, income, healthy lifestyle, and dependents, but did not include a quality attribute. Lancsar et al. (2011) elicited preferences for QALY gains to patients described in terms age at disease onset, age at death, and potential quality of life lost without treatment, but did not consider quality and survival as distinct elements, as both were captured by the QALY. The structure of these elicitations is consistent with the zero condition: in the absence of any life year or QALY gains, the value of changes in the other levels is zero. The attributes included in the current DCE and CSPC elicitations, however, theoretically allowed for improvements in quality over an unchanged life expectancy,¹⁴ and so changes in attribute levels still have value even in the absence of life year gains and the zero condition is not applicable. For this reason, a strictly multiplicative value function was excluded. Interactions with life year gains were included in potential value functions, though, on the grounds that some or even many respondents may view life extension as the primary objective of healthcare.

Linear and design-coded parameters were also tested within the different value functions. Whereas a linear parameter has a single coefficient, implying that the change in utility for a given change in attribute level is constant, design-coded parameter can have multiple coefficients, allowing for non-linear effects over the range of the attribute (Hensher et al. 2005). For example, the utility associated with a three level dummy coded parameter, with level 2 as the reference level, is given by:

$$V = \beta_0 + \beta_1 D1 + \beta_2 D3 \quad (8.1)$$

Where β_1 and β_2 represent the marginal utility associated with levels 1 and 3, respectively, relative to the omitted reference level, level 2, and β_0 is the utility associated with the reference level.

¹⁴ In practice, however, the smallest individual life year gain in the experimental design was 1 year.

Dummy coded parameters are straightforward to code and interpret, but as is clear from the specification above, the utility of the reference level, β_0 , is perfectly confounded with the intercept, also known as the alternative-specific constant (ASC) (Hensher et al. 2005). Confounding is a particular problem when a discrete choice task is defined relative to a fixed comparator or in terms of labelled alternatives, as it is impossible to distinguish the default utility associated with the fixed comparator or labelled alternative from the utility associated with the reference level of the dummy coded attributes. In such cases, effects coding is strongly recommended (Bech & Gyrd-Hansen 2005; Hensher et al. 2005; Louviere et al. 2000b). Like dummy coding, effects coding creates $L-1$ design variables with an excluded reference level. Effects coding for a three-level categorical attribute is shown below, again with level 2 as the reference level:

| | E1 | E3 |
|----|----|----|
| L1 | 1 | 0 |
| L2 | -1 | -1 |
| L3 | 0 | 1 |

In this example, the utility of the excluded reference level 2 is $\beta_0 + \beta_{E1}(-1) + \beta_{E3}(-1)$, or $\beta_0 - (\beta_{E1} + \beta_{E3})$, and can be estimated independently of the ASC (Hensher et al. 2005). However, the interpretation of effects coding is less straightforward than dummy coding as the coefficient on an effects coded parameter represents the deviation of the ‘level mean utility’ from ‘overall mean utility’, which not necessarily the same as the difference from the reference level, particularly in a non-orthogonal experimental design (Hosmer & Lemeshow 2000). Furthermore, confounding is much less of a problem in generic experimental designs, where there is no expectation of a default preference for one alternative or the other as in labelled designs. Indeed, generic designs often assume *a priori* that the ASC is non-significant and exclude it from the choice model (Bech & Gyrd-Hansen 2005). Given the generic design used here, and the challenge of interpreting effects coded parameters, dummy coded parameters were felt to be sufficient for allowing for non-linear effects.

The alternative models and value function specifications were compared on the basis Akaike's information criterion with a correction for finite sample sizes (AICc) and Schwarz's Bayesian information criterion (BIC) (Burnham & Anderson 2004; Magidson & Vermunt 2004). These criteria guide model selection by weighing the trade-off between model fit and parsimony; specifically, the potential for bias stemming from too few parameters, and the potential for imprecision or spurious results stemming from an over-specified model. Both criteria penalise the log-likelihood function (LL) by a factor based on the sample size (n) and the number of parameters (k) in the model: AICc is defined as $-2LL + 2k + 2k(k+1)/(n-k-1)$, which converges to $-2LL + 2k$ in large samples, and BIC is defined as $-2LL + k \cdot \log(n)$. A smaller criterion value is preferred in both cases. Both criteria are commonly used in selecting between discrete choice models, but because BIC applies a larger parameter penalty in reasonably large samples, it tends to favour parsimonious models more strongly than AICc (Swait 2007).

The DCE statistical models were estimated using LIMDEP 9.0/NLOGIT 4.0. Consistent with Hosmer and Lemeshow's (2000) suggestion of relaxing the threshold of statistical significance in order to allow for the broadest possible inclusion of explanatory parameters, a significance threshold of 0.10 was adopted and p-values were not adjusted for multiple comparisons. Dummy coded parameters were excluded from the parsimonious specifications only if all levels were insignificant (Hensher et al. 2005). Robust clustered standard errors for coefficient estimates were calculated using the 'sandwich estimator' (Freedman 2006). Estimates of compensating variation and their associated confidence intervals were calculated in LIMDEP 9.0/NLOGIT 4.0 using the delta method, based on coefficient means and covariances derived from the regression models (Oehlert 1992). A significance threshold of 0.05 was adopted for all other analyses, and p-values were adjusted for multiple simultaneous comparisons using Hommel's method (Shaffer 1995; Wright 1992).

8.1.1 Agent vs. public preferences

A secondary objective of the analysis was to test for heterogeneity between the preferences of self-identified agents and those of the general public.

To test for the effect of agent status on attribute preferences given the limited number of agents who participated in the survey, a classical approach was used, interacting each attribute with agent status (Morey & Greer Rossmann 2003). The baseline value function was based on simple main effects with agent interactions, but a more complex value function with life year gain interactions was also tested. If the interactions between a specific attribute and agent status was found to be significant, the difference in compensating variation between the general population and agents would be calculated and taken as significant if the 95 percent confidence interval around the difference in CV between agents and the general public did not cross zero (Schenker & Gentleman 2001).

8.2 Modelling individual heterogeneity

Each respondent to the survey contributed multiple responses over a series of choice tasks. The simplest approach to analysing such panel data is the 'pooled model,' which assumes that preferences are homogeneous across all individuals (Baltagi 2008). However, if unobserved factors influence the choices made by an individual, particularly as a result of random taste variation or unobserved heterogeneity, these responses will tend to be correlated and treating them as independent observations will reduce the realism of the model and can lead to biased regression estimates (Hole 2008; Glasgow 2001). Random taste variation arises when unobserved individual characteristics influence how the observed characteristics of an alternative affect choice. That is, individuals with the same observed characteristics may place different weights on different aspects of a choice, leading to correlation in the utility of alternatives within a particular choice task. Unobserved heterogeneity arises when an individual's choices depend on unobserved characteristics, leading to correlation in the utility of alternatives between different choice tasks (Glasgow 2001). For example, having children, or experience with a particular disease, may exert an unobserved influence on respondents' choices over a series of choice tasks. Other sources of individual heterogeneity in choice tasks could include response heterogeneity, where respondents utilise response scales differently, perceptual heterogeneity, where respondents differ in their perception of the attributes in a task, and form

heterogeneity, where respondents apply different decision rules in evaluating the alternatives in a task (Desarbo et al. 1997).

Because an individual's personal characteristics are constant across their responses, it is impossible to estimate the effect of these characteristics on the probability of choice. The classical approach to incorporating heterogeneity into a choice model, therefore, has been to interact attributes with individual characteristics. For example, a price attribute may be interacted with gender to determine if males are more sensitive to price than females. As noted in section 4.5, though, this approach has the drawback of assuming that heterogeneity is strictly deterministic, and that everyone with the same observed characteristics must share the same preferences (Boxall & Adamowicz 2002; Morey & Greer Rossmann 2003).

As alternatives, latent class models and random effects or random parameters models allow for more realistic representations of individual heterogeneity in the context of discrete choice experiments. In a conditional logit model of discrete choice (McFadden 1974),

$$\Pr(j)_{it}|\beta_i = \frac{e^{\beta_i x_{itj}}}{\sum_{j=1}^J e^{\beta_i x_{itj}}} \quad (8.2)$$

Where $\Pr(j)_{it}|\beta_i$ is the probability of individual i choosing alternative j in task t given a vector of preference weight β_i , and a vector of attribute levels x_{itj} , individual heterogeneity in preferences can be represented as:

$$\beta_i = \beta + \delta z_i + e_i \quad (8.3)$$

Where β is the mean population preference weight, Δz_i is a vector of individual characteristics and associated coefficients, and e_i is a stochastic individual effect (Hole 2008). Broadly speaking, if the components of e_i are assumed to be continuous and assigned a subjective distribution, this leads to a random effects model. If the components of e_i are assumed to be discrete, this leads to a latent class model (Greene & Hensher 2003; Hole 2008).

Latent class models assume that there are two or more 'classes' or groups underlying the data, which share unobserved (latent) characteristics that affect choice. Critically, preferences are assumed to differ between classes, but to be

homogeneous within classes (Greene & Hensher 2003). As membership in a particular class (c) is a function of latent characteristics, it must be estimated probabilistically, most often using a conventional multinomial logit model:

$$\Pr(C = c | z_i) = \frac{e^{\delta_c z_i}}{\sum_{c=1}^C e^{\delta_c z_i}} \quad (8.4)$$

Where $\Pr(C=c | z_i)$ is the probability of individual i being in class c given a vector of characteristics z_i , $\sum_{c=1}^C \Pr(c) = 1$, and $\delta_c z_i$ is a vector of observed individual characteristics and associated coefficients (Hernández Alava et al. 2012; Provencher et al. 2002). If δ_c is zero, membership in a particular class does not depend on observed characteristics and the likelihood of belonging to any particular class is constant across individuals (Hole 2008).

As the probability of individual i choosing alternative j is conditional upon class membership, choice and class membership must be estimated simultaneously:

$$\Pr(j)_{it} = \left[\frac{e^{\beta_i x_{it} + e_i}}{\sum_{j=1}^J e^{\beta_i x_{it} + e_i}} \right] \left[\frac{e^{\delta_c z_i}}{\sum_{c=1}^C e^{\delta_c z_i}} \right] \quad (8.5)$$

Such a model allows the characteristics of the alternatives and the characteristics of the individual to jointly explain choice behaviour, by weighting the probability of choice by the probability of membership in a discrete number of classes (Ben-Akiva et al. 1997; Boxall & Adamowicz 2002).

In contrast to the latent class model, a random parameters model integrates the probability of choice over all possible values of individual taste and requires subjective assumptions about the distribution of these tastes (Boxall & Adamowicz 2002). In this approach, β_i or e_i is assumed to be a random variable with a subjectively specified distribution which can be interpreted as random variation in individual preferences (individual heterogeneity) or an error term that introduces correlation among the utility of different alternatives (random taste variation) (Amaya-Amaya et al. 2008; Morey & Greer Rossmann 2003). Wedel et al. (1999) noted that the subjective assignment of a distribution has the advantage of being able to force a random parameters model to conform to an underlying theory of behaviour (e.g. by constraining a particular parameter to a positive distribution), as well as facilitating the estimation of individual-level

parameters. They also pointed out that a latent class model cannot fully account for heterogeneity if preferences are in fact continuous. In such cases, latent classes are an artificial partition of continuous preferences and the assumption of homogeneity within those classes is unrealistic. Finally, random parameter models hold the individual random effect, e_i , constant, inducing correlation across an individual's choices, whereas latent class models assume that each choice is an independent draw from a discrete distribution (Greene & Hensher 2003; Shen 2009). However, the subjectivity of the random parameter distributions is also a limitation, as there is little formal theory to guide the selection of which parameters should be specified as random and which distribution to choose. Results are likely to be sensitive to the choice of distribution, and misspecifying a distribution (e.g. constraining a distribution to positive values when in fact there is density on both sides of zero) can lead to biased results (Amaya-Amaya et al. 2008; Greene & Hensher 2003; Hole 2008; Wedel et al. 1999). A continuous distribution of preferences can also be more difficult to interpret than a small number of distinct latent classes (Boxall & Adamowicz 2002; Wedel et al. 1999).

Greene and Hensher (2003) suggested that a latent class model can be thought of as a non-parametric approximation of the continuous random parameters model that avoids the problem of specifying which parameters are in fact random and their distributions. However, latent class models pose the analogous challenge of correctly specifying the number of classes. Like the specification of random distributions in the mixed logit, there is little theory to guide this specification. In practice, classes are generally added so long as the additional class is associated with a decrease in the BIC (Boxall & Adamowicz 2002; Hernández Alava et al. 2012), but this arbitrary approach to specifying the number of latent class offsets to some degree the non-parametric advantage of a latent class model (Greene & Hensher 2003; Hole 2008).

Boxall and Adamowicz (2002) regard the latent class model as a balance between the perfect homogeneity of a pooled model, where each individual is assumed to have identical preferences, and the perfect heterogeneity of a random parameters model, where each individual can be thought of as their own individual latent class. In this light, they characterised the difference between a

random parameters and a latent class model as the difference between *incorporating* heterogeneity and *explaining* heterogeneity. With a random parameters model, individual preferences differ only because each individual is an independent draw from a specified random distribution (Morey & Greer Rossmann 2003), while in a latent class model, individual preferences can be explicitly related to latent or observed characteristics as well as observed choices. Furthermore, in contrast to the classical interaction approach to incorporating heterogeneity, which deterministically divides individuals into groups based solely on observed characteristics, a latent class model assigns individuals to classes probabilistically, allowing for different preferences among individuals with the same observed characteristics. This ability to explain individual heterogeneity as a function of observed and unobserved individual characteristics is the key advantage of the latent class model relative to the classical interaction or random parameters approaches. Although it is almost certainly a simplification to assume homogeneous preferences within latent classes, such simplification greatly enhances the interpretability and salience of the estimates. Latent class models do not appear to have been used previously to analyse stated preferences in a healthcare context, although they have been used in marketing and transportation applications (Ramaswamy & Cohen 2007; Greene & Hensher 2003). Given its potential advantages, a latent class multinomial logit model was tested alongside the simple pooled multinomial logit in identifying a preferred modelling approach.

Within a latent class approach, the ability of the preferred specification to distinguish between the latent classes was assessed in terms of relative entropy (E), or relative classification certainty, defined as:

$$E = 1 - \frac{-\sum_{i=1}^N \sum_{c=1}^C \Pr(c)_{ic} \cdot \log(\Pr(c)_{ic})}{N \cdot \log(C)} \quad (8.6)$$

Where N is the number of individuals in the data, C is the total number of classes in the model, and $\Pr(c)_{ic}$ is the probability of individual i being a member of class c (Dias & Vermunt 2006). Relative entropy is measured on a $[0, 1]$ scale, with values toward 1 suggesting highly stable classifications with clear distinction between classes, and values toward 0 suggesting highly unstable classifications with no clear distinctions.

8.2.1 Relating individual characteristics to latent class membership

The significance of individual characteristics in predicting latent class membership was assessed with a linear logit model, transforming the predicted probability of each individual's membership in class c to the logit scale, and estimating a linear model with dummy-coded parameters for agent status, university or college graduate, gender, and 'fast completer,' along with categorical age group. Although education was not collected as part of the agent questionnaires, it was assumed that all agents had graduated college or university. Given the perfect collinearity between agent status and the graduation flag induced by this assumption, the model was also specified without agent status to exclude its potentially confounding effect. The relative effect of each factor (f) on the overall probability membership in class c [$\Pr(c)$] was calculated as $\beta_f(1 - \Pr(c)) \cdot \Pr(c)$ (Gujarati 1988).

As Clark and Muthén (2009) note in discussing approaches to relating individual covariates to latent class membership, the probability regression approach is superior to deterministically assigning class membership on the basis of the highest probability as it allows for differences in probability between individuals. Specifically, a deterministic approach does not account for the fact that one individual may have a 51 percent probability of class membership while another may have a 99 percent probability of membership; both individuals would be assigned a deterministic weight of 1.0. However, they also note that in treating the probability of class membership as an observation rather than a probabilistic estimate, the probability regression approach may under-estimate the error and potentially over-estimate statistical significance. One way to compensate for this over-estimation of significance is to adopt a more rigorous threshold for statistical significance. However, the magnitude of such under-estimation is negatively related to entropy: as entropy approaches 1.0 (perfect classification certainty), the error associated with treating class membership as an observation approaches zero. In a model with reasonable entropy, the error in significance is not overly problematic. As such, the same 0.10 significance threshold used elsewhere in the analysis was used to assess the statistical significance of the p-values on the individual covariates after adjusting for multiple simultaneous comparisons.

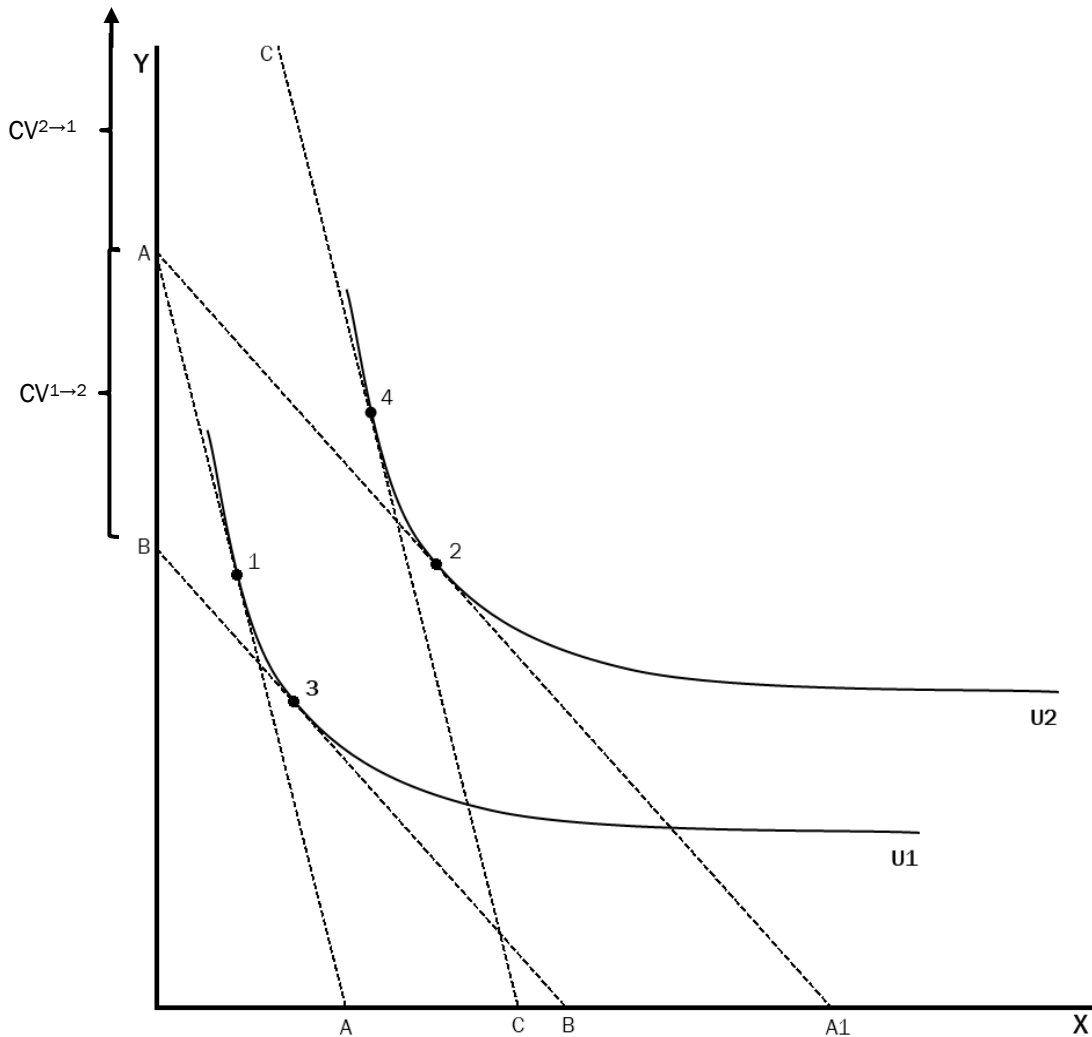
8.3 Estimating welfare effects using compensating variation

The coefficients from the DCE statistical model represented the change in systematic utility given a 1-unit change in an attribute. However, the interpretation of these coefficients is complicated by the fact that many were measured on different scales. Age, for example, was measured in years, while severity was measured on a 0 to 1 quality scale, and the number of patients treated was measured in terms of persons. To transform these marginal utility estimates to a common scale, the analysis of the pilot survey used marginal rates of substitution (MRS), calculated as the ratio of the coefficients on a particular attribute to the coefficient on individual life years gained. This represents the rate at which respondents would be willing to trade-off individual life year gains for a 1-unit change in another attribute. However, this interpretation of MRS only holds when the value function is strictly additive, and for a marginal change in single attribute (Lancsar et al. 2007).

Compensating variation (CV) is conceptually similar to MRS, and indeed if the only change in the 'state of the world' is a 1-level increase or decrease in a single attribute, CV and MRS are identical (Ryan 2004; Silva 2004). However, the advantage of CV is that it can also accommodate discrete changes in multiple attributes, as well as multiplicative interaction terms (Small & Rosen 1981; Lancsar & Savage 2004). This means that CV can value changes in entire scenarios, rather than just a change in a single attribute. CV is measured in terms of the amount of some valued good that an individual would theoretically be willing to sacrifice in order to secure that change. Specifically, it measures how much of that good – the numeraire – could be taken away from an individual following a change so as to leave them at the same level of well-being as before the change (Feldman & Serrano 2006). This is illustrated in Figure 8.1 below.

In this figure, an individual has an initial allocation of goods X and Y shown by point 1 on the budget constraint shown by the dashed line AA , and tangential to the indifference curve U_1 . If, as the result of a policy change, good X becomes relatively less costly and shifts the budget constraint outward to AA_1 , the individual will move to point 2, tangential to the more preferred indifference curve U_2 . This move from point 1 to point 2 is a combination of a 'substitution

Figure 8.1: Illustrating compensating variation



effect' and an 'income effect'. The substitution effect, allowing for a change in prices but holding income constant, moves the utility maximising allocation from point 1 to point 3 on the original indifference curve U_1 . The income effect, allowing for a change in income but holding prices constant, moves the utility maximising point from point 3 to point 4, on the new budget constraint AA_1 and tangential to the more preferred indifference curve U_2 . Together, the two effects combine to move the utility maximising point to point 2. The welfare effect of this policy change, taking Y as the numeraire, can be estimated by the vertical distance between the new budget line, AA_1 , and a hypothetical budget line, BB , parallel to budget line AA_1 and tangential to the original indifference curve at point 3. This distance, $CV^{1 \rightarrow 2}$, is the amount of the numeraire that could be taken away from the individual, given the new implicit prices of X and Y , so as

to leave him exactly as well off as before the change. This can be interpreted as an individual's maximum willingness-to-sacrifice in order to secure a change from point 1 to point 4 (Feldman & Serrano 2006; Zerbe & Dively 1994).

Conversely, if the individual was initially at point 2 and a policy change increases the relative price of X , shifting the budget constraint from $AA1$ to AA , the individual would move to point 1, tangent to the less preferred indifference curve $U1$. The welfare effect associated with this move can be estimated, as above, by the vertical distance between the new budget constraint, AA , and a hypothetical, parallel budget constraint, CC , tangential to the original indifference curve at point 4. In the case of a move to a less preferred point, this distance, $CV^{2 \rightarrow 1}$, represents the minimum amount of the numeraire that the individual would be willing-to-accept in order to agree to the change. However, as illustrated in Figure 8.1 the individual's minimum willingness-to-accept in compensation for a move from point 2 to point 1 ($CV^{2 \rightarrow 1}$) is much larger than his maximum willingness-to-pay to secure a move from point 1 to point 2 ($CV^{1 \rightarrow 2}$), despite the fact that he is moving between the same two points (Feldman & Serrano 2006).

This apparent inconsistency in the welfare effect of a move between points 1 and 2 is driven by the interaction between the income and substitution effects, as the preferred level of X depends on the current level of the numeraire, Y . This can be resolved, though, by assuming that individual utilities are 'quasilinear'. Under this assumption, Y is considered to be a special good that enters every individual's utility function additively. The numeraire can be any good, but is most intuitively understood as wealth or income. A quasilinear utility function can be written as:

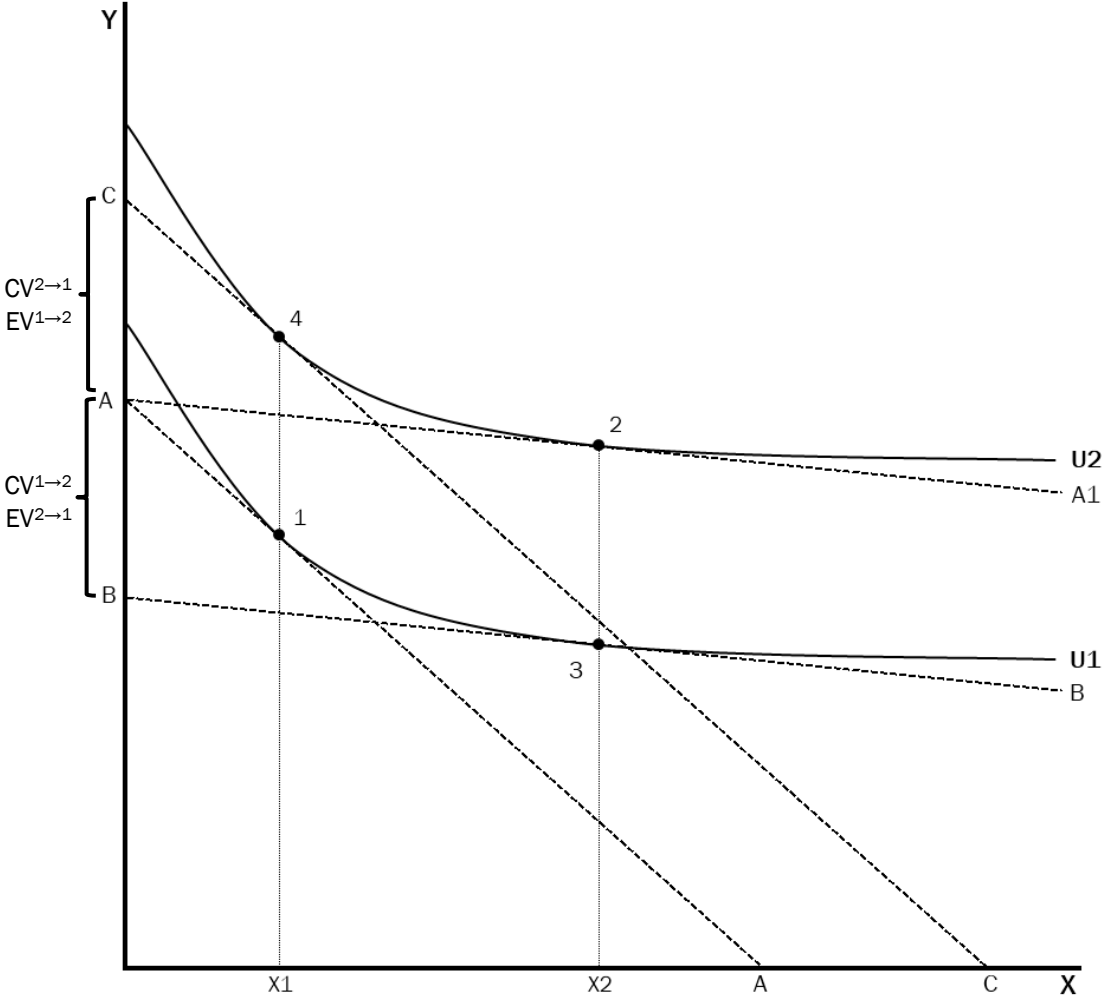
$$U_i = v_i(X_i) + \mu Y_i \quad (8.7)$$

Where individual utility (U_i) is some function of X , plus the amount of the numeraire, Y , weighted by μ , the constant marginal utility of Y . In this form each individual's indifference curves are parallel, with a shape given by $v_i(X_i)$ and a vertical shift given by ΔY_i , with the implication that the preferred level of X does not depend on Y (Feldman & Serrano 2006). The estimation of

compensating variation under an assumption of quasilinear utility functions is illustrated in Figure 8.2.

In this figure, as in Figure 8.1 earlier, an individual is initially at point 1 and moves to point 2 as the result of a decrease in the price of X , shifting the budget constraint from AA to $AA1$. Again, this move is a combination of substitution and income effects as a result of the change in price and the implicit change in income, respectively. The substitution effect, holding income constant, results in a shift from point 1 to point 3 on the initial indifference curve ($U1$) and an increase in the preferred level of X from $X1$ to $X2$. The income effect, holding prices constant, results in a shift from point 3 on the initial indifference curve to point 2 on the more preferred indifference curve ($U2$).

Figure 8.2: Compensating variation with quasilinear utility



However, under an assumption of quasilinear utility, the income effect shown in Figure 8.2 does not induce an additional change in X .

As earlier, the CV associated with a move from point 1 to point 2, based on the prices at point 2, is given by the vertical distance between AA1 and BB, and the CV associated with moving from point 2 back to point 1, based on the prices at point 1, is given by the vertical distance between AA and CC. However, unlike in Figure 8.1, the CV associated with the two moves is equivalent. By removing income effects from the quasilinear utility function, CV is not dependent upon the initial starting point, and the welfare effect, in terms of Y , is consistent regardless of the direction of change. This consistency equates compensating variation with the related concept of equivalent variation (EV). Whereas CV is a measure of the welfare effect based on the new prices, EV is a measure of welfare effect based on the original prices (Feldman & Serrano 2006). For a move from point 1 to the more preferred point 2, EV is a measure of how much of the numeraire an individual would be willing-to-accept in order to forego the move, while for a move from point 2 to the less preferred point 1, EV is a measure of how much of the numeraire an individual would be willing-to-pay in order to prevent the change. As shown in Figure 8.2, under an assumption of quasilinear preferences the willingness-to-pay to secure a move from point 1 to point 2 ($CV^{1 \rightarrow 2}$) is equivalent to the willingness-to-pay to avoid a move from point 2 to point 1 ($EV^{2 \rightarrow 1}$). This symmetry between CV and EV means that the initial position is arbitrary, and that it is possible to consistently, and arguably more intuitively, interpret welfare effects in terms of the willingness-to-pay to secure a move to a more preferred level, or to avoid a move to a less preferred level.

In the context of the elicitation here, quasilinear utility implies that the preferred level of a particular attribute does not depend on the number of individual life years gained, similar to the utility independence condition commonly assumed to underlie the QALY model. This condition holds that preferences for a particular health state are independent of its duration (Pliskin et al. 1980; Miyamoto & Eraker 1988). This is undoubtedly a simplification, as there are reasons other than income effects for a divergence between CV and EV. These include endowment effects, or the idea that individuals value losses more

highly than gains (Kahneman & Tversky 1979), and moral property rights or intrinsic values, which may make more reluctant to accept compensation for a loss than to sacrifice for a gain (Boyce et al. 1992; Shogren et al. 1994). Likewise, Tsuchiya and Dolan (2005) showed that the utility independence assumption may not always hold, and that individuals' preferences for a health state may indeed depend upon its duration. However, the estimation of compensating variation under an assumption of quasilinear utility is consistent with conventional stated preference methods (Jedidi & Zhang 2002; Lancsar & Savage 2004; Lancsar et al. 2007) and has been applied in a recent societal preference elicitation of distributive preferences in healthcare (Baker et al. 2010; Lancsar et al. 2011). Note that this assumption does not imply that respondents would not be willing to sacrifice life year gains in order to secure a more preferred level of an attribute, or that respondents would not be willing to sacrifice other attributes in order to secure greater individual life years gains.

Based on these assumptions, compensating variation, or the change in welfare associated with a change in attribute levels was calculated in the context of a 'state of the world model' (Small & Rosen 1981; Ryan 2004; Silva 2004) as:

$$CV_{a:LYg} = \frac{1}{\beta_{LYg}} [v^0 - v^1] \quad (8.8)$$

Where β_{LYg} is the coefficient on the numeraire, life years gained, or the constant marginal utility of one additional life year gained, and v^0 and v^1 are the scenario utilities before and after a change in one or more attribute levels, respectively. In the case of a move to a *more* preferred scenario, CV will be negative (life year gains must be taken away to return utility to the pre-change level), while in the case of a move to a *less* preferred scenario, CV will be positive (life years must be added to return utility to the pre-change level). The magnitude of the CV estimates for different scenarios where the only change was a 1-level move away from the baseline level was taken to represent the relative strength of preferences for the new levels relative to the baseline level. Note that these potential life year gains were assumed to accrue to other individuals in society, not to the respondent.

The estimate of compensating variation associated with each attribute was used to test the null hypothesis that changes in attribute levels other than individual life years gained would have no impact on welfare, and that respondents would not be willing to sacrifice individual life year gains for equity or distributive justice goals. The use of aggregate, rather than individual, life year gains as the numeraire was considered, but ruled out on the grounds that this would imply a default preference for aggregate life year gains that may not hold. Statistically significant welfare effects were taken as a rejection of the null hypothesis for that attribute.

8.3.1 Scenario rankings

The compensating variation results, which considered marginal preferences holding all other attributes constant, were supplemented by ranking each scenario in the experimental design by its predicted utility, allowing for all attributes to vary simultaneously. DCE scenario utilities were calculated by weighting the attribute levels in each scenario by the attribute coefficients derived from the statistical model. Spearman's rho was calculated to provide a sense of the strength and direction of association between each attribute level and a scenario's relative ranking. As the scenarios were ranked by descending utility, a negative correlation coefficient implies that the relative rank of a scenario improved as an attribute level increased, while a positive correlation coefficient implies that relative rank worsened as an attribute level increased.

The probability of choice for each DCE scenario was calculated relative to a reference scenario with all attributes at their middle (baseline) level, although this scenario was not actually shown to respondents in the choice tasks. The probability of choosing each scenario over the reference scenario was calculated in the context of a conventional multinomial logit model (McFadden 1974):

$$\Pr(i|i, ref) = \frac{e^{\beta_i x_i}}{e^{\beta_i x_i} + e^{\beta_{ref} x_{ref}}} \quad (8.9)$$

Where i was a specific scenario, ref was the reference scenario, and β and x were vectors of attribute coefficients and levels, respectively.

8.4 DCE results

A number of alternative models and functional forms were tested, including pooled and latent class multinomial logit (MNL) models, and strictly additive main effects or main effects with life year gains interactions value functions. The different model and value specifications are shown in Appendix 8.1, ranked by improving log likelihood, AICc and BIC. Only specifications that were associated with an improvement over the previous in terms of at least one of these criteria are shown.

The additive main effects MNL pre-specified at the experimental design stage had the worst fit relative to the other specifications tested. The interaction between final health state and life years gained included in the pre-specified value function was not significant in the additive main effects MNL, but replacing it with an interaction between initial and final health state improved the fit by all information criteria, as did the inclusion of life year interactions with each of the main effects. The main effects remained significant after the introduction of life year interactions, suggesting that respondents may have derived value from allocating resources on the basis of these attributes, independent of the number of life years gained. A parsimonious version of this additive interaction model was associated with an insignificant decrease in log-likelihood and improvements in AICc and BIC. Dummy coded main effects also improved model fit, suggesting non-linearity in preferences across these terms. Despite the penalties associated with the substantial increase in the number of parameters, the latent class model was preferred to the pooled logit on the basis of AICc and BIC, indicative of significant unobserved heterogeneity in preferences. Overall, a parsimonious version of the 3-class latent class logit with continuous main effects and life year gains interactions was preferred by BIC, while a 2-class, dummy coded main effects specification with continuous life year gains main effects and interactions was preferred on the basis of log likelihood and AICc. Although a 3-class dummy coded specification was associated with further improvement in fit by the information criteria, it had very high standard errors in one class. Hole (2008), in discussing the trade-off between model fit and the precision of the parameters, suggested that fewer classes with more precise parameters are generally preferred. A parsimonious

version of the 2-class dummy coded specification, excluding the non-significant interaction between life year gains and final health state, did not converge.

8.4.1 Overall DCE results

Although the more parsimonious 3-class continuous specification was preferred by BIC, the 2-class dummy coded specification had the advantage of allowing for non-linear preferences over the levels presented in the survey, and was preferred in terms of log-likelihood and AICc. The results of this model, weighting the class-specific coefficients by the individual probability of class membership, are shown in Appendix 8.2. The intercept, or alternative specific constant (ASC) was not significant in the overall results, as expected in a generic design. Most of the other coefficients were significant at the 0.10 threshold and moved in the directions anticipated by the empirical ethics review. The insignificant coefficients in these overall results, specifically the dummy on the lower level of total patients treated, and the interactions between life years gained and final health state and total patients treated, were significant in one of the two latent classes and were therefore retained in the overall result. The use of life years gained as the numeraire in estimating compensating variation appeared justified, as the model showed that the marginal utility of an additional life year gained was positive and highly significant ($\beta_{LYg}=0.28$, $p < 0.001$), suggesting that individual life year gains were indeed valued by respondents.

The use of non-linear dummy coded parameters in modelling the overall results appeared to have only limited justification based on the results of Wald tests, shown below in Table 8.2. If preferences were linear over an attribute, the negative slope coefficient at one end of the range would offset the positive slope coefficient at the other end of the range, and the sum of the two coefficients would not be significantly different than zero. The Wald tests showed that the sum of the slope coefficients on the high and low dummy-coded parameters were not significantly different than zero for age and initial health state, while final health state was only significant at a 0.10 threshold.

Table 8.2: Wald tests of non-linearity in dummy-coded parameters

| Attribute levels | Difference | Std. Error | Diff/Std. err | Adj. p-value | Sig |
|--------------------------|------------|------------|---------------|--------------|-----|
| Age(10) + Age(70) | 0.41 | 0.34 | 1.21 | 0.45 | |
| U0(0.1) + U0(0.9) | -0.24 | 0.37 | -0.63 | 0.53 | |
| LE(1m) + LE(10yrs) | -0.79 | 0.19 | -4.08 | <0.001 | *** |
| U1(0.1) + U1(0.9) | -1.02 | 0.44 | -2.30 | 0.06 | + |
| nPats(100) + nPats(5000) | 1.37 | 0.55 | 2.48 | 0.04 | * |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

U0=initial health state; LE=life expectancy; U1=final health state; nPats=number of patients treated

The compensating variation (CV) associated with an upward or downward change in the level of each attribute, relative to a baseline state with all attributes at their middle level, are also shown graphically in Figure 8.3, and detailed in Table 8.3. To clearly illustrate which attribute levels were more preferred and less preferred relative to the reference level, the y-axis was reversed to show more preferred scenarios (negative CV) above zero, and less preferred scenarios (positive CV) below zero.

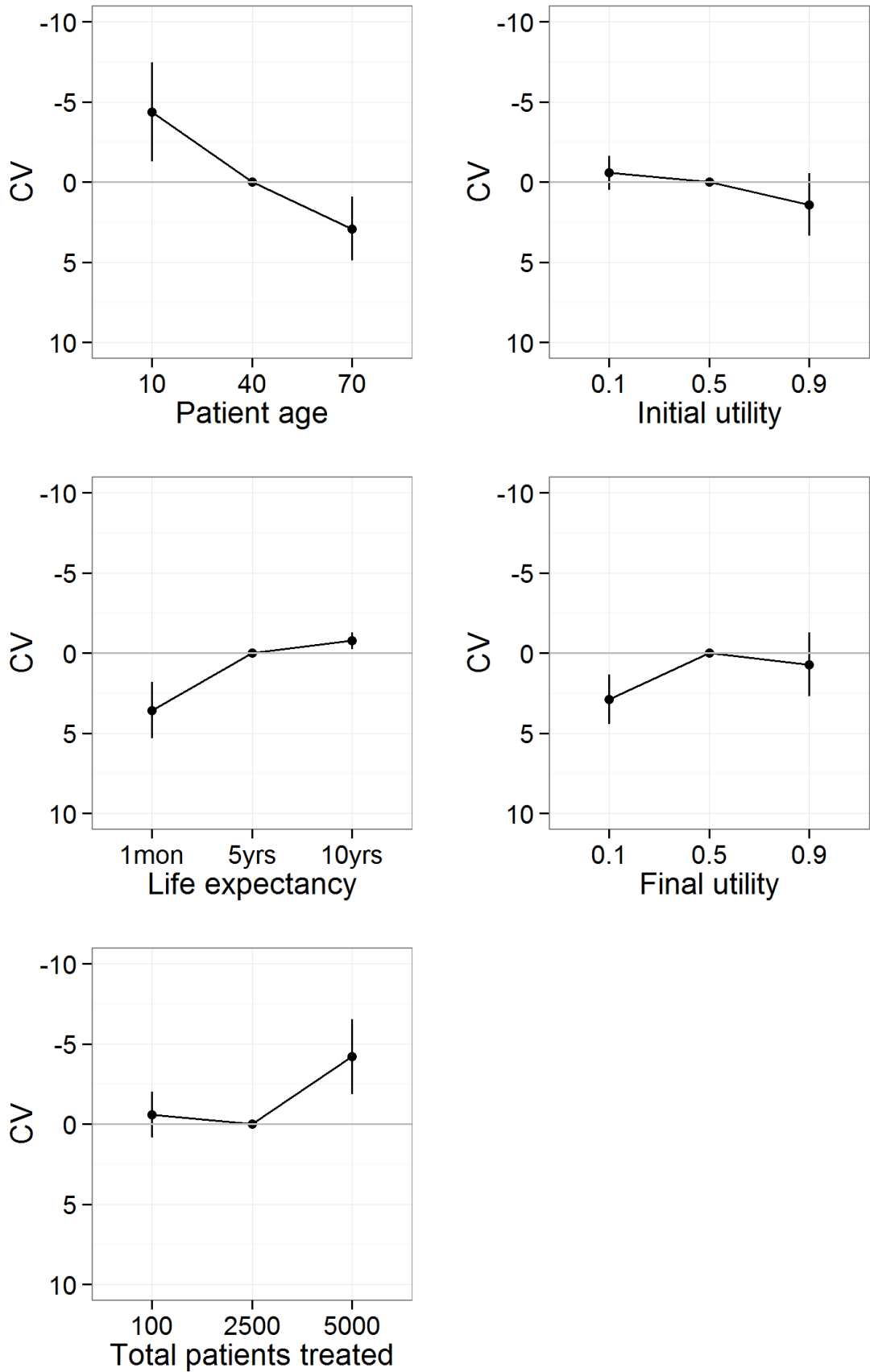
Table 8.3: DCE compensating variations by change in attribute levels

| Attributes | Attribute levels | CV (95% CI), Baseline → Low | CV (95% CI), Baseline → High |
|------------------------|-----------------------|--------------------------------|---------------------------------|
| Patient age | 10y/o - 40y/o - 70y/o | -4.36 (-7.45, -1.26) | 2.91 (0.91, 4.91) |
| Initial health state | 0.1 - 0.5 - 0.9 | -0.57 (-1.63, 0.48) | 1.41 (-0.55, 3.36) |
| Life expectancy | 1m - 5yrs - 10yrs | 3.57 (1.82, 5.32) | -0.77 (-1.30, -0.25) |
| Final health state | 0.1 - 0.5 - 0.9 | 2.88 (1.34, 4.43) | 0.71 (-1.27, 2.69) |
| Total patients treated | 100 - 2500 - 5000 | -0.60 (-2.03, 0.83) | -4.20 (-6.55, -1.86) |

CV=compensating variation; 95% CI = 95% confidence interval. CVs are for a change away from the baseline (middle) level, holding all other attributes at their baseline level. Statistically significant CVs are shown in bold.

Confidence intervals that did not cross zero were taken to be statistically significant. Negative CVs indicated a positive welfare effect (i.e. a quantity of the numeraire could be taken away following a change in attribute levels and leave respondents at least as well-off as before the change), and positive CVs indicated a negative welfare effect (i.e. a quantity of the numeraire would have to be added following a change in attribute levels to leave respondents at least as well-off as before the change).

Figure 8.3: DCE compensating variations by attribute



The statistically significant CVs associated with the upper and lower levels of patient age suggested that there were positive welfare effects associated with prioritising 10-year-old patients, and negative welfare effects associated with prioritising 70 year-old patients, although this effect was weaker than in the younger age group. Contrary to the expectation of a preference for prioritising more severe patients suggested by the empirical ethics review, there were no significant effects over initial health state and negative welfare effects associated with prioritising patients with the shortest untreated life expectancy. There was a small but statistically significant welfare gain associated with prioritising patients with the longest initial life expectancy. There was also a significant welfare loss associated with prioritising patients that would finish in the worst final health state after treatment, but no significant effect associated with patients that ended up in the best final health state. Finally, there was a significant welfare gain associated with treating 5000 over 2500 patients, but no significant welfare loss associated with treating 100 rather than 2500 patients.

Overall, the greatest welfare gains were associated with prioritising 10-year-old patients over 40-year-old patients, and treating an additional 2500 patients over the baseline scenario. Conversely, the greatest welfare losses were associated with giving priority to patients with the shortest life expectancy, the oldest age or the worst final health state. The greatest absolute difference in CV between the high and low levels of an attribute, taken as an indicator of relative importance, was over patient age ($\Delta CV=7.27$, 95% CI: 2.53, 12.01), followed by individual life years gained ($\Delta CV=5.43$, 95% CI: 2.32, 8.54).

8.4.2 DCE scenario rankings

The utility associated with each DCE scenario was calculated by weighting the attribute levels in each choice alternative by the overall finite mixture model coefficients shown in Appendix 8.2. The 11 choice tasks in each of the two design blocks presented 22 different choice sets, for a total of 44 scenarios. The two choice sets (4 scenarios) that were re-presented in repeated task of each block were excluded, as was one of the choice sets (2 scenarios) from the test of dominance in the second design block, as this set was identical in both blocks, leaving a total of 38 scenarios to be ranked. A reference scenario,

with all attributes at their middle (baseline) level, was included as a comparator, although this scenario was not actually shown to respondents.

These scenarios, ranked from most to least preferred in terms of their predicted utility and probability of choice relative to the reference scenario, are presented in Table 8.4, along with Spearman correlation coefficients showing the association between attribute levels and scenario rank. As a reminder, the scenarios were ranked by descending utility, so a negative correlation coefficient implies that the relative rank of a scenario *improved* as an attribute level increased, while a positive correlation coefficient implies that rank *worsened* as an attribute level increased.

Table 8.4: DCE scenario rankings by predicted utility and probability of choice

| Rank | Age | UO | LE | U1 | LYg | nPats | Ind. QALYs | Agg. QALYs | Utility | Prob. of choice |
|-------------|-----------|------------|----------|------------|----------|-------------|-------------|-------------|-------------|-----------------|
| 1 | 10 | 0.5 | 5 | 0.5 | 10 | 5000 | 5.00 | 25000 | 5.19 | 90.5% |
| 2 | 40 | 0.1 | 0.083 | 0.9 | 10 | 2500 | 9.07 | 22666 | 5.01 | 88.9% |
| 2 | 40 | 0.1 | 0.083 | 0.9 | 10 | 2500 | 9.07 | 22666 | 5.01 | 88.9% |
| 4 | 10 | 0.1 | 10 | 0.5 | 5 | 100 | 6.50 | 650 | 4.81 | 86.8% |
| 5 | 10 | 0.1 | 0.083 | 0.5 | 10 | 100 | 5.03 | 503 | 4.81 | 86.7% |
| 6 | 10 | 0.5 | 10 | 0.5 | 1 | 2500 | 0.50 | 1250 | 4.54 | 83.3% |
| 7 | 10 | 0.5 | 5 | 0.9 | 10 | 100 | 11.00 | 1100 | 4.47 | 82.3% |
| 8 | 70 | 0.5 | 5 | 0.9 | 10 | 5000 | 11.00 | 55000 | 4.37 | 80.8% |
| 9 | 40 | 0.5 | 10 | 0.5 | 5 | 5000 | 2.50 | 12500 | 4.36 | 80.7% |
| 10 | 40 | 0.5 | 10 | 0.5 | 10 | 2500 | 5.00 | 12500 | 3.84 | 71.2% |
| 11 | 10 | 0.5 | 0.083 | 0.1 | 10 | 5000 | 0.97 | 4834 | 3.62 | 66.5% |
| 12 | 70 | 0.1 | 0.083 | 0.9 | 5 | 5000 | 4.57 | 22832 | 3.56 | 65.2% |
| 13 | 40 | 0.1 | 10 | 0.1 | 10 | 5000 | 1.00 | 5000 | 3.50 | 63.8% |
| 14 | 10 | 0.5 | 0.083 | 0.1 | 1 | 5000 | 0.07 | 334 | 3.26 | 58.1% |
| 15 | 40 | 0.5 | 10 | 0.9 | 5 | 100 | 8.50 | 850 | 3.23 | 57.4% |
| 16 | 40 | 0.1 | 5 | 0.5 | 1 | 5000 | 2.50 | 12500 | 3.15 | 55.4% |
| 17 | 70 | 0.1 | 5 | 0.1 | 10 | 100 | 1.00 | 100 | 2.95 | 50.4% |
| Ref. | 40 | 0.5 | 5 | 0.5 | 5 | 2500 | 2.50 | 6250 | 2.93 | 50.0% |
| 18 | 10 | 0.9 | 0.083 | 0.9 | 1 | 5000 | 0.90 | 4500 | 2.81 | 46.9% |
| 19 | 40 | 0.9 | 0.083 | 0.5 | 10 | 5000 | 4.97 | 24834 | 2.74 | 45.2% |
| 20 | 10 | 0.5 | 0.083 | 0.1 | 10 | 2500 | 0.97 | 2417 | 2.72 | 44.7% |
| 21 | 70 | 0.9 | 10 | 0.5 | 10 | 2500 | 1.00 | 2500 | 2.63 | 42.4% |
| 22 | 10 | 0.1 | 10 | 0.1 | 5 | 2500 | 0.50 | 1250 | 2.38 | 36.6% |
| 23 | 10 | 0.9 | 5 | 0.9 | 5 | 100 | 4.50 | 450 | 2.38 | 36.6% |
| 23 | 70 | 0.1 | 5 | 0.5 | 5 | 2500 | 4.50 | 11250 | 2.27 | 34.1% |
| 25 | 10 | 0.9 | 5 | 0.9 | 5 | 2500 | 4.50 | 11250 | 2.17 | 31.9% |

| | | | | | | | | | | |
|-----------|----|-----|-------|-----|----|------|------|-------|-------|-------|
| 26 | 40 | 0.5 | 5 | 0.9 | 1 | 2500 | 2.90 | 7250 | 2.11 | 30.4% |
| 27 | 40 | 0.9 | 0.083 | 0.1 | 5 | 2500 | 0.43 | 1084 | 2.04 | 29.0% |
| 28 | 40 | 0.1 | 0.083 | 0.5 | 5 | 2500 | 2.53 | 6333 | 2.02 | 28.6% |
| 29 | 70 | 0.9 | 10 | 0.9 | 10 | 5000 | 9.00 | 45000 | 1.86 | 25.4% |
| 30 | 10 | 0.1 | 10 | 0.1 | 1 | 2500 | 0.10 | 250 | 1.83 | 24.9% |
| 31 | 70 | 0.1 | 10 | 0.1 | 5 | 5000 | 0.50 | 2500 | 1.48 | 19.0% |
| 32 | 40 | 0.9 | 10 | 0.9 | 1 | 100 | 0.90 | 90 | 1.40 | 17.7% |
| 33 | 40 | 0.5 | 0.083 | 0.1 | 5 | 100 | 0.47 | 47 | 1.25 | 15.7% |
| 34 | 70 | 0.5 | 0.083 | 0.5 | 5 | 100 | 2.50 | 250 | 1.25 | 15.7% |
| 35 | 70 | 0.5 | 10 | 0.9 | 1 | 100 | 4.90 | 490 | 1.18 | 14.8% |
| 36 | 40 | 0.9 | 0.083 | 0.5 | 1 | 2500 | 0.47 | 1167 | 1.15 | 14.4% |
| 37 | 40 | 0.1 | 0.083 | 0.5 | 1 | 100 | 0.53 | 53 | 0.55 | 8.4% |
| 38 | 40 | 0.1 | 5 | 0.1 | 1 | 100 | 0.10 | 10 | -0.47 | 3.2% |

| | | | | | | | | | | |
|-------------------|-------------|-------------|-------------|--------------|--------------|--------------|--------------|--------------|--|--|
| Rank corr. | 0.34 | 0.21 | 0.06 | -0.21 | -0.53 | -0.22 | -0.55 | -0.44 | | |
|-------------------|-------------|-------------|-------------|--------------|--------------|--------------|--------------|--------------|--|--|

Age=patient age; U0=initial health state; LE=initial life expectancy; U1=final health state; nPats=patients treated; LYg=individual life years gained; Ind. QALYs=QALYs gained per patient; Agg. QALYs=Aggregate QALYs (individual QALYs weighted by total patients treated); Utility=Predicted utility from DCE choice model; Prob. of choice=Probability of choosing a particular scenario compared to the reference scenario. The reference scenario is shown in **bold**.

The correlation coefficients suggested that respondents valued individual health gains, with larger individual QALY gains and individual life year gains having moderate to strong associations with better scenario rankings. Larger aggregate QALY gains and better final health states were also associated with better scenario rankings, while increasing patient age and initial health state were associated with poorer rankings. Each of the top five scenarios had individual QALY gains in the top 20 percent across all scenarios, and three of the top five scenarios had aggregate QALY gains in the top 10% across all scenarios. Likewise, seven of the bottom ten scenarios had aggregate QALY gains in the bottom 20 percent across all scenarios. However, four scenarios among the top ten, all presenting 10 year old patients, had aggregate QALY gains well below the median, and the two scenarios with the largest and the second largest aggregate QALY gains, in both cases accruing to 70 year old patients, were ranked 8th and 29th out of the 38 scenarios. Although QALY gains appeared to be strongly associated with higher rankings, scenarios with relatively small aggregate QALY gains were often ranked favourably when these gains accrued

to the youngest patients, while relatively large QALY gains to older patients were less favourably ranked.

The predicted probability of choice of the most preferred scenario suggests that 91 percent of respondents would be expected choose that scenario over the reference, while only 3 percent of respondents would be expected to choose the least preferred alternative over the reference. By definition, the reference scenario had a 50 percent probability of choice relative to itself, which can be interpreted as indifference, or an equal probability of choice between two identical alternatives.

In order to control for the strong effect of age and more clearly illustrate how the other attributes interacted to drive choice, the scenarios are re-presented in Table 8.5 ordered by utility and choice probability within each age level as a form of two-way sensitivity analysis. Note that because the experimental design was not perfectly orthogonal, the number of scenarios in each age stratum is not equal.

Table 8.5: Age-stratified DCE scenario rankings by predicted utility and probability of choice

| Overall rank | Rank within age | U0 | LE | U1 | LYg | nPats | Ind. QALYs | Agg. QALYs | Utility | Pr(Choice) |
|---------------|-----------------|-----|-------|-----|-----|-------|------------|------------|---------|------------|
| Age 10 | | | | | | | | | | |
| 1 | 1 | 0.5 | 5 | 0.5 | 10 | 5000 | 5.00 | 25,000 | 5.19 | 90.5% |
| 4 | 2 | 0.1 | 10 | 0.5 | 5 | 100 | 6.50 | 650 | 4.81 | 86.8% |
| 5 | 3 | 0.1 | 0.083 | 0.5 | 10 | 100 | 5.03 | 503 | 4.81 | 86.7% |
| 6 | 4 | 0.5 | 10 | 0.5 | 1 | 2500 | 0.50 | 1,250 | 4.54 | 83.3% |
| 7 | 5 | 0.5 | 5 | 0.9 | 10 | 100 | 11.00 | 1,100 | 4.47 | 82.3% |
| 11 | 6 | 0.5 | 0.083 | 0.1 | 10 | 5000 | 0.97 | 4,834 | 3.62 | 66.5% |
| 14 | 7 | 0.5 | 0.083 | 0.1 | 1 | 5000 | 0.07 | 334 | 3.26 | 58.1% |
| 18 | 8 | 0.9 | 0.083 | 0.9 | 1 | 5000 | 0.90 | 4,500 | 2.81 | 46.9% |
| 20 | 9 | 0.5 | 0.083 | 0.1 | 10 | 2500 | 0.97 | 2,417 | 2.72 | 44.7% |
| 22 | 10 | 0.1 | 10 | 0.1 | 5 | 2500 | 0.50 | 1,250 | 2.38 | 36.6% |
| 23 | 11 | 0.9 | 5 | 0.9 | 5 | 100 | 4.50 | 450 | 2.38 | 36.6% |
| 25 | 12 | 0.9 | 5 | 0.9 | 5 | 2500 | 4.50 | 11,250 | 2.17 | 31.9% |
| 30 | 13 | 0.1 | 10 | 0.1 | 1 | 2500 | 0.10 | 250 | 1.83 | 24.9% |
| Age 40 | | | | | | | | | | |
| 2 | 1 | 0.1 | 0.083 | 0.9 | 10 | 2500 | 9.07 | 22,666 | 5.01 | 88.9% |
| 9 | 2 | 0.5 | 10 | 0.5 | 5 | 5000 | 2.50 | 12,500 | 4.36 | 80.7% |

| | | | | | | | | | | |
|-------------|-------------|------------|----------|------------|----------|-------------|-------------|--------------|-------------|--------------|
| 10 | 3 | 0.5 | 10 | 0.5 | 10 | 2500 | 5.00 | 12,500 | 3.84 | 71.2% |
| 13 | 4 | 0.1 | 10 | 0.1 | 10 | 5000 | 1.00 | 5,000 | 3.50 | 63.8% |
| 15 | 5 | 0.5 | 10 | 0.9 | 5 | 100 | 8.50 | 850 | 3.23 | 57.4% |
| 16 | 6 | 0.1 | 5 | 0.5 | 1 | 5000 | 2.50 | 12,500 | 3.15 | 55.4% |
| Ref. | Ref. | 0.5 | 5 | 0.5 | 5 | 2500 | 2.50 | 6,250 | 2.93 | 50.0% |
| 19 | 7 | 0.9 | 0.083 | 0.5 | 10 | 5000 | 4.97 | 24,834 | 2.74 | 45.2% |
| 26 | 8 | 0.5 | 5 | 0.9 | 1 | 2500 | 2.90 | 7,250 | 2.11 | 30.4% |
| 27 | 9 | 0.9 | 0.083 | 0.1 | 5 | 2500 | 0.43 | 1,084 | 2.04 | 29.0% |
| 28 | 10 | 0.1 | 0.083 | 0.5 | 5 | 2500 | 2.53 | 6,333 | 2.02 | 28.6% |
| 32 | 11 | 0.9 | 10 | 0.9 | 1 | 100 | 0.90 | 90 | 1.40 | 17.7% |
| 33 | 12 | 0.5 | 0.083 | 0.1 | 5 | 100 | 0.47 | 47 | 1.25 | 15.7% |
| 36 | 13 | 0.9 | 0.083 | 0.5 | 1 | 2500 | 0.47 | 1,167 | 1.15 | 14.4% |
| 37 | 14 | 0.1 | 0.083 | 0.5 | 1 | 100 | 0.53 | 53 | 0.55 | 8.4% |
| 38 | 15 | 0.1 | 5 | 0.1 | 1 | 100 | 0.10 | 10 | -0.47 | 3.2% |

Age 70

| | | | | | | | | | | |
|----|---|-----|-------|-----|----|------|-------|--------|------|-------|
| 8 | 1 | 0.5 | 5 | 0.9 | 10 | 5000 | 11.00 | 55,000 | 4.37 | 80.8% |
| 12 | 2 | 0.1 | 0.083 | 0.9 | 5 | 5000 | 4.57 | 22,832 | 3.56 | 65.2% |
| 17 | 3 | 0.1 | 5 | 0.1 | 10 | 100 | 1.00 | 100 | 2.95 | 50.4% |
| 21 | 4 | 0.9 | 10 | 0.5 | 10 | 2500 | 1.00 | 2,500 | 2.63 | 42.4% |
| 23 | 5 | 0.1 | 5 | 0.5 | 5 | 2500 | 4.50 | 11,250 | 2.27 | 34.1% |
| 29 | 6 | 0.9 | 10 | 0.9 | 10 | 5000 | 9.00 | 45,000 | 1.86 | 25.4% |
| 31 | 7 | 0.1 | 10 | 0.1 | 5 | 5000 | 0.50 | 2,500 | 1.48 | 19.0% |
| 34 | 8 | 0.5 | 0.083 | 0.5 | 5 | 100 | 2.50 | 250 | 1.25 | 15.7% |
| 35 | 9 | 0.5 | 10 | 0.9 | 1 | 100 | 4.90 | 490 | 1.18 | 14.8% |

Age=patient age; U0=initial health state; LE=initial life expectancy; U1=final health state; nPats=patients treated; LYg=individual life years gained; Ind. QALYs=QALYs gained per patient; Agg. QALYs=Aggregate QALYs (individual QALYs weighted by total patients treated); Utility=Predicted utility from DCE choice model; Pr(Choice)=Probability of choosing a particular scenario compared to the reference scenario. The reference scenario is shown in **bold**.

Consistent with the overall results, the age-stratified results appeared to emphasise the importance of survival gains and aggregate QALYs, as the most highly ranked scenario within each age strata had the highest level of individual life year gains as well as substantial aggregate QALY gains. However, there also appeared to be an offsetting preference against individuals in the best initial health state, as scenarios associated with some of the greatest aggregate QALY gains were ranked relatively poorly when they accrued to patients in the best initial health state. The absolute gain in health-related utility appeared less important than survival gains, as a number of the highly ranked scenarios within each age strata were associated with no change between the initial and final health states.

The scenarios were also stratified by initial life expectancy (not shown) to explore its relationship with individual life year gains. The empirical ethics review suggested that respondents might be indifferent over the number of life years gained in scenarios where patients faced imminent death on the grounds that any gain would be valuable, because as Harris (1985) argues, it is all the time they have left. Contrary to this hypothesis, however, scenarios with larger individual life year gains were consistently ranked more favourably than scenarios with smaller life year gains when patients had a life expectancy of only 1 month. A similar pattern was found in scenarios where life expectancy was 5 years, but there was no clear preference over life year gains when initial life expectancy was 10 years.

8.4.3 DCE results by latent class

Overall, the probability of being in a particular latent class was approximately equal, as there was a 48 percent probability of belonging to class 1 and a 52 percent probability of belonging to class 2. As shown in Figure 8.4, however, the individual probability of membership had a bimodal distribution, with peaks at very high and very low probabilities of membership, suggesting a clear distinction between classes at the individual level. This was supported by the estimate of the relative entropy, which measured the model's ability to

distinguish between

latent classes. The

relative entropy of the

two-class model was

0.67 on a 0-1 scale,

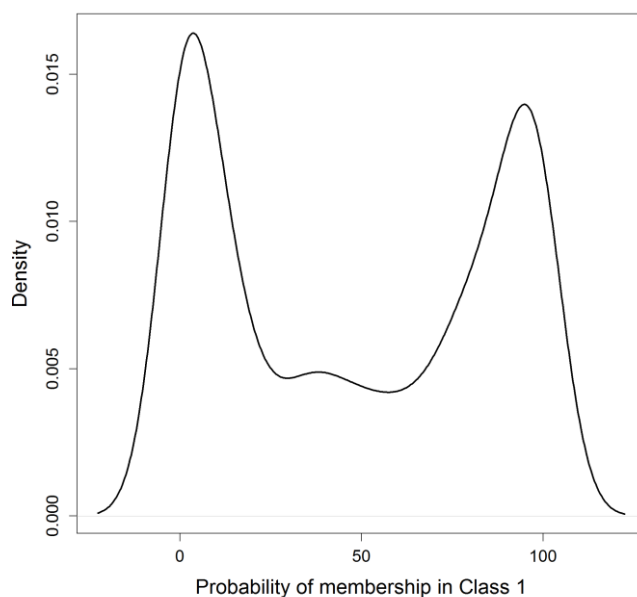
suggesting a moderate

ability to distinguish

between the latent

classes.

Figure 8.4: Latent class 1 membership probability density



The logit-transformed probability regression model found that university or college graduation, gender and age group were not statistically significant predictors of the probability of latent class membership, but agent and ‘fast completer’ status were significant at a 0.10 threshold. After excluding the insignificant parameters and re-estimating the model, agent status was associated with a statistically significant 33 percent relative reduction in the probability of membership in class 1 (adjusted- $p=0.04$). From an overall probability of belonging to latent class 1 of 48 percent, the probability of an agent belonging to class 1 was 32 percent, with a corresponding 68 percent probability of belonging to class 2, suggesting that agents were twice as likely to belong to class 2 as class 1. Fast completion was associated with a 25 percent reduction in the probability of membership in class 1, but this reduction failed to meet a 0.10 significance threshold (adjusted- $p=0.12$). A model excluding agent status, specified to avoid possible confounding with education, found that none of the remaining factors were significant at a 0.10 threshold.

The latent class coefficients shown in Appendix 8.3 indicated that the majority of coefficients were significant at a 0.10 threshold in both classes, although the standard errors were notably larger in class 1 than in class 2. The alternative specific constant in class 1 was not significant, although the constant in class 2 was significant and positive, suggesting some *a priori* preference for Alternative B (the right-hand side alternative) among these respondents. Several other coefficients were notable for the difference in sign and magnitude between the two classes. For example, the signs on the coefficients on the lowest and highest levels of patient age, as well as the age-life years gained interaction term, were reversed between class 1 and class 2. In addition, the size of the coefficients on the dummy-coded age parameters was substantially different. However, as the coefficients on the dummy-coded main effects and the age-life year interaction term moved in opposite directions in the two classes, it was difficult to anticipate the net effect of a change in age on expected utility and compensating variation. The coefficients also showed that use of life years gained as the numeraire in the compensating variation calculations was justified by its positive and significant coefficient in both classes. The difference in the

marginal utility of an additional life year gained between classes was not significant (difference=0.01, p=0.92).

Wald tests on the sum of the slope coefficients on the high and low levels of the dummy-coded parameters, shown in Table 8.6, were much more suggestive of non-linearity in the main effects by latent class than in the overall results. Interestingly, although the magnitude of the non-linearities was generally smaller in class 2 than class 1, they were also more strongly significant.

Table 8.6: Wald tests of non-linearity in dummy-coded parameters, by latent class

| Attribute | Difference | Std. Error | β /Std. err | Adj. p-value | Sig |
|--|------------|------------|-------------------|--------------|-----|
| Age(10) + Age(70), Class 1 | 0.37 | 0.70 | 0.54 | 0.59 | |
| U0(0.1) + U0(0.9), Class 1 | -2.65 | 0.72 | -3.68 | <0.001 | *** |
| LE(1m) + LE(10yrs), Class 1 | -0.91 | 0.36 | -2.56 | 0.03 | * |
| U1(0.1) + U1(0.9), Class 1 | -2.41 | 0.86 | -2.81 | 0.02 | * |
| nPats(100) + nPats(5000), Class 1 | 3.22 | 1.05 | 3.07 | 0.01 | * |
| Age(10) + Age(70), Class 2 | 0.44 | 0.07 | 6.33 | <0.001 | *** |
| U0(0.1) + U0(0.9), Class 2 | 1.99 | 0.10 | 20.01 | <0.001 | *** |
| LE(1m) + LE(10yrs), Class 2 | -0.67 | 0.07 | -9.15 | <0.001 | *** |
| U1(0.1) + U1(0.9), Class 2 | 0.27 | 0.13 | 2.11 | 0.07 | + |
| nPats(100)+ nPats(5000), Class 2 | -0.35 | 0.09 | -4.06 | <0.001 | *** |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+' | | | | | |

U0=initial health state; LE=life expectancy; U1=final health state; nPats=number of patients treated

The compensating variations by attribute within each class, and the net differences between classes, are shown in Table 8.7. They suggest that the strength and direction of preferences were generally consistent in the two classes, although there were significant differences in the direction of preference for initial health states, and the best final health state. There were also significant differences in the strength of preference for the number of patients treated.

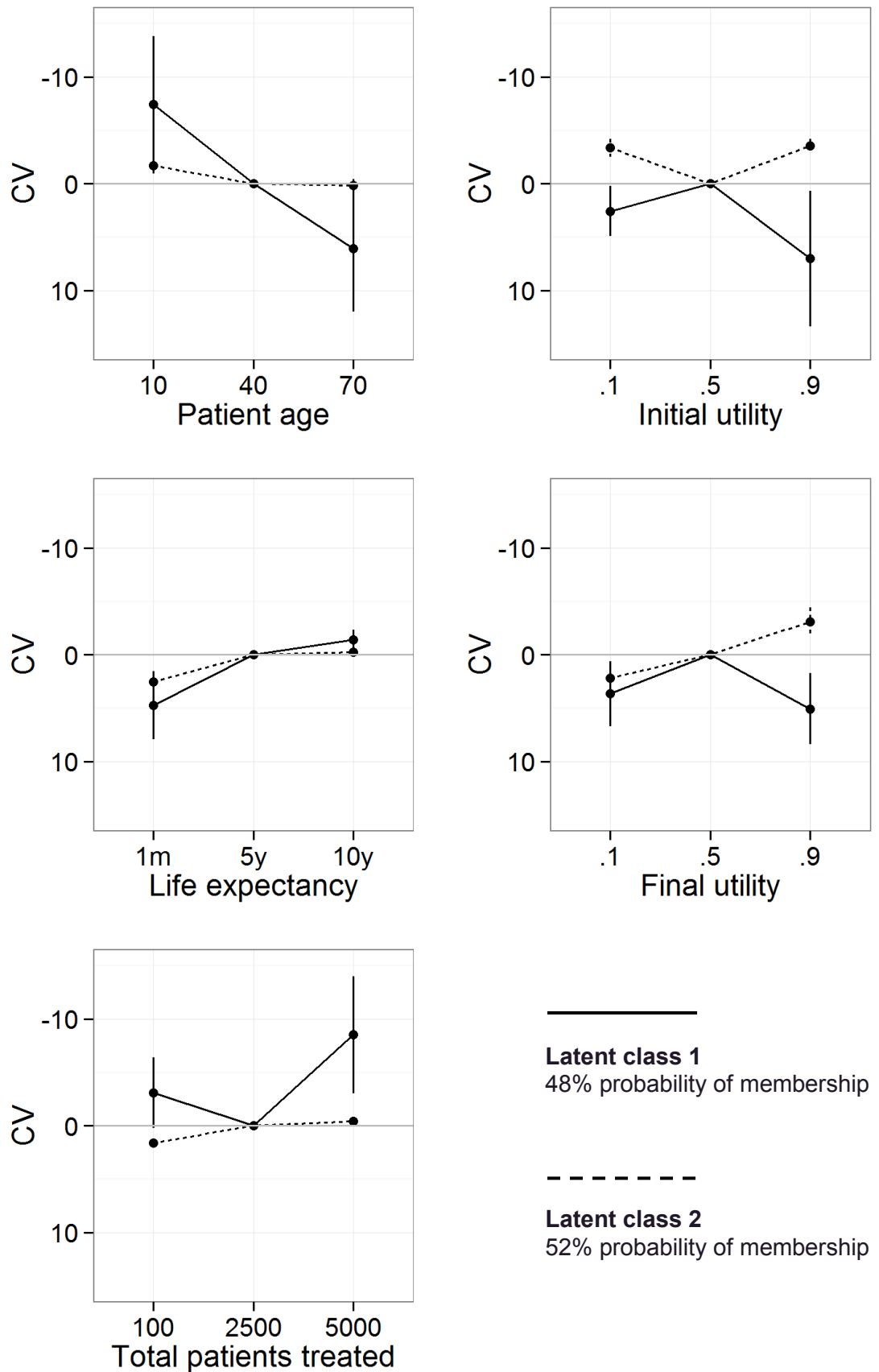
Table 8.7: Compensating variations and differences between latent classes by attribute change

| Attribute change | CV (95% CI) Class 1 | CV (95% CI) Class 2 | Difference (95% CI) Class 1 - Class 2 |
|-------------------------------------|--------------------------|-------------------------|--|
| Patient age, 40 → 10 | -7.41 (-13.80, -1.02) | -1.67 (-2.38, -0.97) | -5.74 (-12.18, 0.71) |
| Patient age, 40 → 70 | 6.06 (0.15, 11.97) | 0.14 (-0.39, 0.68) | 5.91 (-0.09, 11.92) |
| Initial health state, 0.5 → 0.1 | 2.58 (0.24, 4.91) | -3.35 (-4.17, -2.52) | 5.92 (3.54, 8.30) |
| Initial health state, 0.5 → 0.9 | 7.02 (0.68, 13.37) | -3.54 (-4.21, -2.86) | 10.56 (4.24, 16.88) |
| Life expectancy, 5yrs → 1mon | 4.72 (1.54, 7.91) | 2.55 (1.93, 3.17) | 2.18 (-1.12, 5.47) |
| Life expectancy, 5yrs → 10yrs | -1.38 (-2.34, -0.43) | -0.23 (-0.50, 0.03) | -1.15 (-2.18, -0.12) |
| Final health state, 0.5 → 0.1 | 3.66 (0.64, 6.69) | 2.20 (1.53, 2.86) | 1.47 (-1.70, 4.63) |
| Final health state, 0.5 → 0.9 | 5.06 (1.72, 8.40) | -3.11 (-4.42, -1.80) | 8.17 (4.67, 11.68) |
| Total patients treated, 2500 → 100 | -3.09 (-6.37, 0.19) | 1.59 (1.26, 1.93) | -4.69 (-7.99, -1.39) |
| Total patients treated, 2500 → 5000 | -8.51 (-13.97, -3.04) | -0.41 (-0.73, -0.09) | -8.10 (-13.57, -2.63) |

Statistically significant differences are shown in **bold**.

Figure 8.5, on the next page, shows that there was a significant and positive welfare effect (negative CV) associated with prioritising patients in the worst initial health state in latent class 2, but a significant and negative welfare effect (positive CV) in latent class 1. A similar opposing pattern was observed for patient groups in the best initial and final health states. There was a negative welfare effect in class 1 associated with patients groups in the best initial health state, but a positive effect in class 2. There was also a negative welfare effect in class 1 associated with patient groups that would end up in the best final health state, and a positive effect in class 2.

Figure 8.5: DCE compensating variation by attribute and latent class



8.4.4 Public vs. agent preferences

The reference model for the DCE choice model with agent interactions was a pooled multinomial logit (MNL) with dummy coded main effects, continuous life year gain interactions and agent interactions (see Appendix 8.4 for a comparison of the alternative value functions, ranked by log-likelihood, AICc and BIC). It was felt that there were not enough agent respondents to justify stratifying them further with a latent class approach. A parsimonious version of the MNL model was preferred by AICc and BIC, and was not significantly worse than the full model by the likelihood ratio test. As in the previous models, a significance threshold of 0.10 was adopted, and dummy coded parameters were only excluded if the entire system of coefficients were insignificant. The results of this model are shown in Appendix 8.5.

Agents appeared to hold more moderate preferences than the general population sample over the high and low levels of initial health state (U0), as the coefficients on the initial health state-agent interactions tended to offset the main effects coefficients. Agents also appeared to hold divergent preferences for the worst final health state (U1). Although the general population coefficient was positive, the coefficient on the agent interaction term was negative and much larger, suggesting a contradictory preference. The compensating variations associated with changes in the initial and final health states are shown below in Table 8.8. These results show that although a move from the baseline to worst (lowest) final health state was associated with negative welfare effects (positive CV) in both groups, the effect was significantly stronger among agents. The mean difference in the effect associated with a move from the baseline to best (highest) initial health state was just significant at a 0.05 threshold. The difference in the welfare effect between the two groups over a move to the best (highest) level of initial health attribute reflects the fact that agents had no statistically significant preference over either state, while the general public had a significant preference for patient groups in the better initial health state.

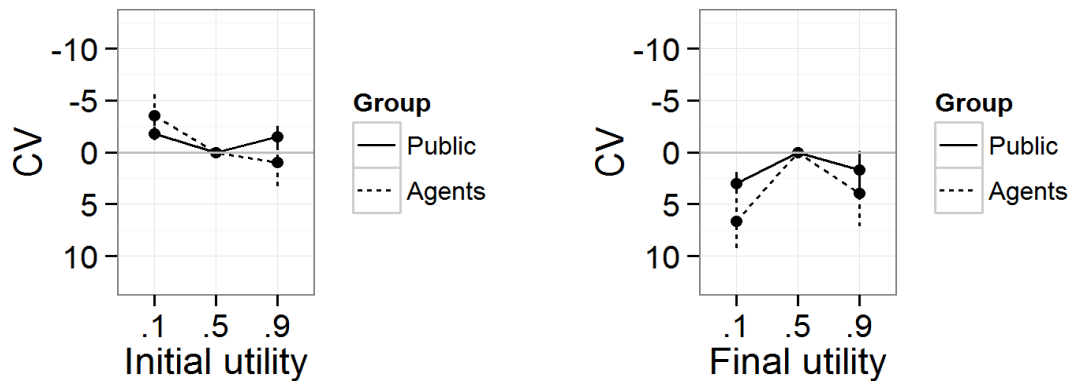
Table 8.8: Compensating variations and differences for agents and the general population

| Attribute change | Compensating variation | Lower 95% CI | Upper 95% CI |
|--|------------------------|--------------|--------------|
| CV, U0 baseline → low, Public | -1.78 | -2.42 | -1.13 |
| CV, U0 baseline → low, Agents | -3.54 | -5.56 | -1.51 |
| Difference, U0 baseline → low, Agents- Public | -1.76 | -3.79 | 0.27 |
| CV, U0 baseline → high, Public | -1.53 | -2.52 | -0.54 |
| CV, U0 baseline → high, Agents | 0.99 | -1.53 | 3.51 |
| Difference, U0 baseline → high, Agents- Public | 2.52 | 0.00 | 5.04 |
| CV, U1 baseline → low, Public | 2.97 | 1.93 | 4.02 |
| CV, U1 baseline → low, Agents | 6.61 | 3.61 | 9.60 |
| Difference, U1 0.5 → 0.1, Agents- Public | 3.63 | 0.70 | 6.57 |
| CV, U1 baseline → high, Public | 1.71 | -0.12 | 3.53 |
| CV, U1 baseline → high, Agents | 3.93 | 0.68 | 7.18 |
| Difference, U1 baseline → high, Agents- Public | 2.22 | -0.56 | 4.99 |

CV=compensating variation; ΔCV =difference in compensating variation ($CV_{agents}-CV_{Public}$). U0=Initial utility; U1=Final utility. Statistically significant differences are shown in **bold**.

The compensating variations for moves between the levels of initial and final health states, by group, are shown graphically in Figure 8.6. They suggest that even for attributes with statistically significant differences in compensating variation, the overall direction of preferences was reasonably consistent in both groups, with the exception of the effect associated with a move from baseline to the best initial health state, where the direction of effect, rather than just the relative strength, was significantly different.

Figure 8.6: Compensating variations for changes in initial and final health state, by group



8.5 Discussion of DCE results

Consistent with the empirical ethics review, the overall results from the DCE suggested that respondents had statistically significant preferences for younger patient groups, larger patient groups, and greater individual life year gains. Despite the significant preferences for larger patient groups and greater individual life year gains, the interaction between these two terms was not significant, suggesting that preferences for these factors were not related to preferences for *aggregate* life year gains. Indeed, this interaction was negative and statistically significant in latent class 1, strongly suggesting diminishing returns to aggregate life years gained, while it was statistically insignificant in class 2. Instead, the preference for larger patient groups appeared to reflect a desire to distribute healthcare benefits as widely as possible. Also, although the interaction between initial and final health state was significant and positive, suggesting a preference for absolute quality gain, the interaction between quality gain and individual life years gained, or, in effect, individual quality-adjusted life years gained, was not significant in either latent class. These overall compensating variation results appeared consistent with the scenario rankings, which also suggested that individual gains were more important to respondents than aggregate gains, and that smaller benefits accruing to preferred patient groups were often preferred over larger gains to less preferred groups.

In contrast to the empirical ethics review of Chapter 3, which found a preference for health gains to individuals in more severe health states, DCE respondents had a preference for patients with longer untreated life expectancies, even after controlling for potential health gains, and no significant preference for patient groups in the most severe initial health state relative to those in better initial states. This non-significant effect appeared to be driven by differences between the latent classes over this attribute: whereas class 2 had a significant preference for prioritising patients in the worst initial health state relative to those in the moderate state, class 1 had a significant aversion to prioritising such patients. This result was mirrored by similarly unexpected preferences for patients in the best initial health state: whereas class 2 had a significant preference for prioritising patients in the best initial health state over those in the moderate state, class 1 had a significant aversion to prioritising such patients. These offsetting preferences led to a statistically insignificant overall result despite statistically significant preferences over initial health state in both classes. This highlights the value of latent class modelling, which allows such heterogeneity to be incorporated, and just as importantly, interpreted.

Respondents had a significant aversion to patient groups that would be in the worst final health state following treatment, but no significant preference for patients in the best final health state relative to patients in a moderate final health state. Again, this result was driven by offsetting differences between classes: although class 2 had a significant preference for patients in the best final health state, class 1 had an even stronger aversion to such patients. This result, though, is not inconsistent with the empirical ethics review, as although there was evidence of a reluctance to allocate resources to patients that would remain in a poor health state following treatment, this did not appear to translate into a preference for patients in the best final health state. It has been suggested that such a pattern may imply a preference for achieving some minimum level of quality in the post-treatment health state rather than maximising the quality of that health state (Schwappach 2002b; Dolan, Cookson 2000).

As described above, latent class 2 had significantly stronger preferences for patients in the worst initial health state and the best final health state, compared to patients in moderate initial and final health states. This appeared to

reflect a greater concern for absolute quality gain relative to class 1. However, confidence intervals around the estimates of compensating variation for latent class 2 were substantially smaller than the corresponding intervals for latent class 1, suggesting that the defining latent characteristic may not be the relative *strength* of preferences, but rather the relative *homogeneity* of preferences. Individuals in class 2 appeared to share a well-formed set of preferences, while preferences in class 1 were consistently more heterogeneous, ranging from very strong to barely significant. Indeed, the latent classes may even reflect the difference between respondents with axiomatically rational preferences (complete, stable and transitive), and those with axiomatically irrational or poorly-formed preferences. It is worth noting in this context that despite the roughly equal overall probabilities of membership in the two classes, agents were statistically much more likely to belong to latent class 2. In light of evidence that respondents may construct their preferences as they progress through a stated preference elicitation (Payne et al. 1992; Ryan 2009; Slovic 1995), and to the extent that agents may be expected to be somewhat more familiar with their preferences over the attributes tested here than the general public, this may lend support the notion that the latent classes reflect differences in the consistency and ‘quality’ of these preferences.

Reinterpreting the DCE latent class results in this light lends support to the notion of a distinction between well-defined versus vaguely-defined preferences. In particular, the very large compensating variations associated with age, initial health state and total patients treated in latent class 1 may reflect non-compensatory decision-making heuristics that favoured younger patients and larger patient groups, and discriminated against those in better final health states, without regard for other attribute levels. In such cases, compensating variation would essentially be infinite, as respondents would theoretically be willing to sacrifice any number of individual life years in order to prioritise their preferred group. If the distinction between the latent classes indeed reflects the quality and consistency of the underlying preferences, it has implications for the role of naive public respondents in societal priority setting, and whose preferences should be accepted as representative. At the extreme, one approach might be to use latent class modelling to identify and exclude individual

respondents with poorly-formed preferences, although efforts to improve respondent's understanding and preference construction would seem to be more in keeping with a democratic or Communitarian approach.

Overall, the DCE results suggest a broadly utilitarian preference, with larger QALY gains tending to be associated with greater expected utility, although respondents were clearly willing to deviate from this rule to prioritise younger or larger patient groups. The corresponding CSPP methods and results will be presented in the next chapter.

Appendix 8.1: Alternative DCE models and value functions, by improving information criteria

| Model and value function | k | LL [Pr(χ^2)] | AICc | BIC |
|---|----|------------------------|--------|--------|
| 1.0) MNL; continuous main effects + U1:LYg interaction (pre-specified at experimental design stage) v = LYg + Age + U0 + LE + U1 + nPats + U1:LYg | 8 | -4031.4 | 8078.8 | 8133.1 |
| 2.0) MNL; continuous main effects + U0:U1 interaction v = LYg + Age + U0 + LE + U1 + nPats + (1-U0):U1 | 8 | -3872.1 | 7760.1 | 7814.4 |
| 3.0) MNL; continuous main effects + LYg interactions + U0:U1 interaction v = LYg + Age + U0 + LE + U1 + nPats + LYg:Age + LYg:U0 + LYg:LE + LYg:U1 + LYg:nPats + (1-U0):U1 | 13 | -3855.6 | 7737.3 | 7825.4 |
| 3.1) Parsimonious MNL; continuous main effects + LYg interactions + U0:U1 interaction v = LYg + Age + U0 + LE + U1 + nPats + LYg:U0 + LYg:LE + LYg:U1 + (1-U0):U1 | 11 | -3860.6 [0.007] | 7739.2 | 7800.3 |
| 4.0) MNL; continuous LYg + dummy-coded main effects + U0:U1 interaction v = LYg + D_AgeL1 + D_AgeL3 + D_U0L1 + D_U0L3 + D_LEL1 + D_LEL3 + D_U1L1 + D_U1L3 + D_nPatsL1 + D_nPatsL3 + LYg:Age + LYg:U0 + LYg:LE + LYg:U1 + LYg:nPats + (1-U0):U1 | 18 | -3846.1 | 7718.3 | 7806.5 |
| 5.0) MNL; continuous LYg + dummy-coded main effects + continuous LYg and U0:U1 interactions v = LYg + D_AgeL1 + D_AgeL3 + D_U0L1 + D_U0L3 + D_LEL1 + D_LEL3 + D_U1L1 + D_U1L3 + D_nPatsL1 + D_nPatsL3 + LYg:Age + LYg:U0 + LYg:LE + LYg:U1 + LYg:nPats + (1-U0):U1 | 18 | -3815.1 | 7666.2 | 7788.2 |
| 5.1) Parsimonious MNL; continuous LYg + dummy-coded main effects + continuous LYg and U0:U1 interactions v = LYg + D_AgeL1 + D_AgeL3 + D_U0L1 + D_U0L3 + D_LEL1 + D_LEL3 + D_U1L1 + D_U1L3 + D_nPatsL1 + D_nPatsL3 + LYg:U0 + LYg:LE + LYg:U1 + (1-U0):U1 | 16 | -3818.0 [0.055] | 7666.0 | 7767.7 |
| 6.0) 2-class LC-MNL; continuous main effects + LYg interactions + U0:U1 interaction v = LYg + Age + U0 + LE + U1 + nPats + LYg:Age + LYg:U0 + LYg:LE + LYg:U1 + LYg:nPats + (1-U0):U1 | 26 | -3763.5 | 7583.2 | 7773.0 |
| 6.1) 3-class LC-MNL; continuous main effects + LYg interactions + U0:U1 interaction v = LYg + Age + U0 + LE + U1 + nPats + LYg:Age + | 39 | -3685.1 | 7454.7 | 7739.1 |

LYg:U0 + LYg:LE + LYg:U1 + LYg:nPats + (1-U0):U1

6.2) Parsimonious 3-class LC-MNL; continuous main effects + LYg interactions + U0:U1 interaction

$v = \text{LYg} + \text{Age} + \text{U0} + \text{LE} + \text{U1} + \text{nPats} + \text{LYg:U0} + \text{LYg:LE} + \text{LYg:U1} + (1-\text{U0}):U1$

33 -3679.5
 [0.082] 7431.5 **7675.3**

7.0) 2-class LC-MNL; continuous LYg + dummy-coded main effects + continuous LYg and U0:U1 interactions

$v = \text{LYg} + \text{D_AgeL1} + \text{D_AgeL3} + \text{D_U0L1} + \text{D_U0L3} + \text{D_LEL1} + \text{D_LEL3} + \text{D_U1L1} + \text{D_U1L3} + \text{D_nPatsL1} + \text{D_nPatsL3} + \text{LYg:Age} + \text{LYg:U0} + \text{LYg:LE} + \text{LYg:U1} + \text{LYg:nPats} + (1-\text{U0}):U1$

37 -3674.4 **7425.2** 7682.6

k =parameters, including alternative specific constant; LL=Log-likelihood; AICc= Akaike information criterion, with correction for finite sample size; BIC=Bayesian information criterion. MNL=multinomial logit; LC-MNL=latent class multinomial logit. Only models and value functions associated with an improvement in LL, AICc or BIC over the previous specification are shown. The overall minimum log-likelihood, AICc and BIC are shown in bold. The p-value of the likelihood ratio [$\text{Pr}(\chi^2)$] is shown for nested models.

Appendix 8.2: Combined latent class model coefficients

| Attribute | Coefficient | Std. Error | Coef of var | β /Std. err | Pr(> z) | Sig |
|---------------|-------------|------------|-------------|-------------------|----------|-----|
| Intercept | -0.13 | 0.18 | 1.385 | -0.72 | 0.4709 | |
| LYg | 0.28 | 0.06 | 0.214 | 4.64 | 0.0000 | *** |
| Age 10 | 1.98 | 0.86 | 0.434 | 2.30 | 0.0213 | * |
| Age 70 | -1.57 | 0.59 | 0.376 | -2.67 | 0.0076 | ** |
| U0 0.1 | -2.33 | 0.76 | 0.326 | -3.07 | 0.0022 | ** |
| U0 0.9 | 2.09 | 0.47 | 0.225 | 4.44 | 0.0000 | *** |
| LE 1m | -1.32 | 0.36 | 0.273 | -3.71 | 0.0002 | *** |
| LE 10yrs | 0.53 | 0.22 | 0.415 | 2.38 | 0.0174 | ** |
| U1 0.1 | 0.85 | 0.26 | 0.306 | 3.28 | 0.0010 | ** |
| U1 0.9 | -1.86 | 0.64 | 0.344 | -2.90 | 0.0037 | ** |
| 100 patients | -0.03 | 0.15 | 5.000 | -0.21 | 0.8338 | |
| 5000 patients | 1.40 | 0.49 | 0.350 | 2.85 | 0.0044 | ** |
| (1-U0):U1 | 8.75 | 2.08 | 0.238 | 4.21 | 0.0000 | *** |
| LYg:Age | 0.05 | 0.03 | 0.600 | 1.73 | 0.0829 | + |
| LYg:U0 | -0.37 | 0.14 | 0.378 | -2.66 | 0.0078 | ** |
| LYg:LE | -0.01 | 0.01 | 1.000 | -2.12 | 0.0338 | * |
| LYg:U1 | -0.04 | 0.06 | 1.500 | -0.78 | 0.4378 | |
| LYg:nPats | -0.02 | 0.01 | 0.500 | -1.43 | 0.1543 | |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+'

Coefficients are based on the latent class coefficients weighted by the individual probabilities of class membership. LYg=individual life year gains; U0=initial utility; LE=initial life expectancy; U1=final utility; nPats=number of patients treated.

Appendix 8.3: Latent class model coefficients, by class

| Attribute | Coefficient | Robust SE | β /Std. err | Pr(> z) | Sig |
|--|-------------|-----------|-------------------|----------|-----|
| Probability, Class 1 | 0.48 | 0.04 | 13.12 | 0.000 | *** |
| Constant | -0.67 | 0.39 | -1.74 | 0.081 | + |
| LYg | 0.28 | 0.13 | 2.57 | 0.027 | * |
| Age 10 | 4.81 | 1.64 | 4.09 | 0.003 | ** |
| Age 70 | -4.43 | 1.08 | -5.06 | 0.000 | *** |
| U0 0.1 | -4.13 | 1.47 | -4.29 | 0.005 | ** |
| U0 0.9 | 1.48 | 0.95 | 2.16 | 0.118 | |
| LE 1m | -2.01 | 0.68 | -4.49 | 0.003 | ** |
| LE 10yrs | 1.10 | 0.46 | 3.27 | 0.017 | ** |
| U1 0.1 | 0.60 | 0.52 | 1.40 | 0.253 | |
| U1 0.9 | -3.01 | 1.26 | -3.34 | 0.017 | ** |
| 100 patients | 0.33 | 0.28 | 0.97 | 0.244 | |
| 5000 patients | 2.89 | 0.97 | 4.78 | 0.003 | ** |
| (1-U0):U1 | 8.10 | 4.10 | 3.03 | 0.048 | * |
| LYg:Age | 0.18 | 0.05 | 4.39 | 0.001 | *** |
| LYg:U0 | -0.90 | 0.26 | -4.46 | 0.001 | *** |
| LYg:LE | -0.03 | 0.01 | -3.04 | 0.016 | * |
| LYg:U1 | -0.01 | 0.12 | -0.04 | 0.962 | |
| LYg:Pats | -0.04 | 0.03 | -2.42 | 0.083 | + |
| Probability, Class 2 | 0.52 | 0.03 | 15.29 | 0.000 | *** |
| Constant | 0.37 | 0.06 | 6.61 | 0.000 | *** |
| LYg | 0.29 | 0.04 | 2.57 | 0.000 | *** |
| Age 10 | -0.63 | 0.14 | 4.09 | 0.000 | ** |
| Age 70 | 1.07 | 0.16 | -5.06 | 0.000 | *** |
| U0 0.1 | -0.67 | 0.12 | -4.29 | 0.000 | *** |
| U0 0.9 | 2.66 | 0.17 | 2.16 | 0.000 | *** |
| LE 1m | -0.68 | 0.08 | -4.49 | 0.000 | *** |
| LE 10yrs | 0.01 | 0.07 | 3.27 | 0.896 | |
| U1 0.1 | 1.07 | 0.11 | 1.40 | 0.000 | *** |
| U1 0.9 | -0.81 | 0.17 | -3.34 | 0.000 | *** |
| 100 patients | -0.36 | 0.11 | 0.97 | 0.001 | ** |
| 5000 patients | 0.01 | 0.11 | 4.78 | 0.895 | |
| (1-U0):U1 | 9.35 | 0.35 | 3.03 | 0.000 | *** |
| LYg:Age | -0.07 | 0.01 | 4.39 | 0.000 | *** |
| LYg:U0 | 0.12 | 0.05 | -4.46 | 0.011 | * |
| LYg:LE | 0.00 | 0.00 | -3.04 | 0.217 | |
| LYg:U1 | -0.08 | 0.03 | -0.04 | 0.010 | * |
| LYg:Pats | 0.01 | 0.01 | -2.42 | 0.122 | |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+' | | | | | |

LYg=life year gains; U0=initial utility; LE=initial life expectancy; U1=final utility; nPats=patients treated.

Appendix 8.4: Alternative DCE public-agent interaction value functions, by improving information criteria

| Attributes | p-values, specification 1 | p-values, specification 2 | p-values, specification 3 |
|-----------------------|---------------------------|---------------------------|---------------------------|
| Constant | 0.150 | 0.199 | |
| LYg | 0.001 | 0.000 | 0.000 |
| D1_Age | 0.027 | 0.004 | 0.000 |
| D3_Age | 0.884 | 0.444 | 0.000 |
| D1_U0 | 0.000 | 0.000 | 0.000 |
| D3_U0 | 0.000 | 0.000 | 0.000 |
| D1_LE0 | 0.000 | 0.000 | 0.000 |
| D3_LE0 | 0.008 | 0.001 | 0.001 |
| D1_U1 | 0.050 | 0.045 | 0.021 |
| D3_U1 | 0.000 | 0.000 | 0.000 |
| D1_nPats | 0.810 | 0.544 | 0.879 |
| D3_nPats | 0.000 | 0.000 | 0.000 |
| (1-U0):U1 | 0.000 | 0.000 | 0.000 |
| LYg:(Age/10) | 0.209 | 0.273 | |
| LYg:U0 | 0.589 | | |
| LYg:LE0 | 0.019 | 0.005 | 0.006 |
| LYg:U1 | 0.000 | 0.000 | 0.000 |
| LYg:(Pats/1000) | 0.952 | | |
| LYg:(1-U0:U1) | 0.000 | 0.000 | 0.000 |
| LYg:Agent | 0.527 | | |
| D1_Age:Agent | 0.894 | | |
| D3_Age:Agent | 0.373 | | |
| D1_U0:Agent | 0.067 | 0.051 | 0.085 |
| D3_U0:Agent | 0.033 | 0.038 | 0.047 |
| D1_LE0:Agent | 0.127 | 0.552 | |
| D3_LE0:Agent | 0.177 | 0.182 | |
| D1_U1:Agent | 0.162 | 0.011 | 0.013 |
| D3_U1:Agent | 0.347 | 0.138 | 0.114 |
| D1_Pats:Agent | 0.687 | | |
| D3_Pats:Agent | 0.831 | | |
| (1-U0:U1):Agent | 0.575 | | |
| LYg:(Age/10):Agent | 0.575 | | |
| LYg:U0:Agent | 0.555 | | |
| LYg:LE0:Agent | 0.540 | | |
| LYg:U1:Agent | 0.718 | | |
| LYg:(Pats/1000):Agent | 0.760 | | |
| LYg:(1-U0:U1):Agent | 0.310 | | |

| Parameters | 37 | 23 | 19 |
|----------------|---------|---------|---------|
| LL | -3777.8 | -3785.7 | -3788.4 |
| Pr(χ^2) | -- | 0.33 | 0.27 |
| AICc | 7634.2 | 7617.5 | 7615.0 |
| BIC | 7795.5 | 7773.4 | 7743.8 |

Specifications are based on a pooled multinomial logit. LL=Log-likelihood; AICc= Akaike information criterion, with correction for finite sample size; BIC=Bayesian information criterion. Only value functions associated with an improvement in LL, AICc or BIC over the previous specification are shown. The overall minimum log-likelihood, AICc and BIC are shown in bold. The p-value of the likelihood ratio Pr(χ^2) is shown relative to the full specification. LYg=individual life year gains; U0=initial utility; LE=initial life expectancy; U1=final utility; nPats=number of patients treated; D1=level 1 dummy; D3=level 3 dummy.

Appendix 8.5: Dummy-coded MNL with agent interactions coefficients

| Attribute | Estimate | Std. Error | t-value | Pr(> t) | Sig |
|---------------|----------|------------|---------|----------|-----|
| LYg | 0.15 | 0.01 | 11.22 | 0.000 | *** |
| Age 10 | 0.66 | 0.06 | 10.98 | 0.000 | *** |
| Age 70 | -0.34 | 0.08 | -4.45 | 0.000 | *** |
| U0 0.1 | -1.11 | 0.11 | -10.27 | 0.000 | *** |
| U0 0.9 | 1.60 | 0.16 | 10.28 | 0.000 | *** |
| LE 1m | -1.16 | 0.10 | -12.14 | 0.000 | *** |
| LE 10yrs | 0.23 | 0.07 | 3.25 | 0.001 | ** |
| U1 0.1 | 0.35 | 0.15 | 2.30 | 0.021 | * |
| U1 0.9 | -1.04 | 0.13 | -8.23 | 0.000 | *** |
| 100 patients | 0.01 | 0.06 | 0.15 | 0.879 | |
| 5000 patients | 0.86 | 0.06 | 13.89 | 0.000 | *** |
| (1-U0):U1 | 3.66 | 0.60 | 6.06 | 0.000 | *** |
| LE:LYg | -0.01 | 0.00 | -2.75 | 0.006 | ** |
| U1:LYg | -0.29 | 0.03 | -9.57 | 0.000 | *** |
| (1-U0):U1:LYg | 0.64 | 0.08 | 8.37 | 0.000 | *** |
| U0 0.1:Agent | 0.26 | 0.15 | 1.73 | 0.085 | + |
| U0 0.9:Agent | -0.37 | 0.19 | -1.98 | 0.047 | * |
| U1 0.1:Agent | -0.53 | 0.22 | -2.48 | 0.013 | * |
| U1 0.9:Agent | -0.33 | 0.21 | -1.58 | 0.114 | |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+'

LYg=individual life year gains; U0=initial utility; LE=initial life expectancy; U1=final utility; nPats=number of patients treated

Chapter 9: Primary CSPC results

As in the previous chapter, the primary objective of the analysis of the CSPC responses was to estimate the relative strength of preferences for the patient and program characteristics identified in the empirical ethics review. The approach used to estimate the strength of these preferences was similar to that taken in the previous chapter: the marginal utility associated with changes in each of the attributes was modelled, and the welfare effects associated with these changes are reported in terms of compensating variations. But whereas the DCE asked respondents to choose one group to prioritise, the CSPC asked respondents to allocate a fixed budget between the two groups. This difference in the response format had implications for how the responses should be modelled, and for how compensating variation should be estimated and interpreted.

This chapter outlines the methods used in modelling and estimating welfare effects, and discuss the results. Section 9.1 describes the specification of a linear CSPC model, allowing for the continuous CSPC response format and the panel nature of the responses. The estimation and interpretation of compensating variation as a measure of welfare effects in light of this response format, including a comparison of public and agent preferences, is discussed in section 9.2. Section 9.3 describes the methods used to rank the CSPC scenarios by relative utility, as the DCE scenarios were in the previous chapter, in order to consider respondent preferences in a more holistic context. As noted, the CSPC has an arguable advantage over DCE in allowing respondents to express a preference for a maximising or equalising distribution of resources or outcomes, independent of the characteristics of the particular choice scenarios. Section 9.4

describes the methods used to identify these specific distributive preferences. Finally, to compare the results of the CSPC in a more direct way to those of the DCE, section 9.5 describes the methods for transforming the CSPC allocations to discrete choices and the non-linear model used to analyse the results. The results of these analyses, including the model coefficients, estimates of welfare effects, scenario rankings, distributive preferences, and a direct comparison of DCE and CSPC discrete choices, are described in section 9.6. Section 9.7 discusses the results.

9.1 Specifying the CSPC model

As with the DCE multinomial logit model, the simplest approach to analysing panel data is the linear ‘pooled model’, which implies that preferences are the same across all individuals (i) and all tasks (t):

$$y_{it} = \alpha + \beta x_t + u_{it} \quad (9.1)$$

Where y_{it} is a continuous response variable, α and β are assumed to be the homogeneous for all individuals and all responses, and u_{it} is a stochastic individual error term with a mean of zero (Croissant & Millo 2008). If there is heterogeneity in the parameters, however, an ‘unobserved effects model’ may be more appropriate:

$$y_{it} = \alpha_{it} + \beta_{it} x_t + \mu_i + \varepsilon_{it} \quad (9.2)$$

Where α and β are assumed to be heterogeneous across respondents (‘individual effect’), across responses (‘time effect’), or both (‘two-way effect’). The unobserved effects model separates the random error term of the pooled model, u_{it} , into two components: an individual-specific component (μ_i), and a stochastic error term (ε_{it}) (Baltagi 2008; Croissant & Millo 2008):

$$u_{it} = \mu_i + \varepsilon_{it} \quad (9.3)$$

Although in practice the individual and the stochastic error terms are not separately identifiable, assumptions about their behaviour lead to fixed or random specifications of the unobserved effects model (Croissant & Millo 2008).

Louviere, Hensher and Swait (2000a) suggest that CSPC is consistent with random utility theory (RUT) and can yield cardinal utility measures if it can be assumed that the differences in the budget allocation reflect differences in latent utility between the two alternatives. Under this interpretation the difference in latent value is theoretically unbounded, even though the observed budget difference ($\Delta Budget = Budget^B - Budget^A$) is bounded by -100 (the entire budget to program A) and +100 (the entire budget to program B). As this implies that the observed budget differences are censored representations of the difference in latent utility, a censored regression, or tobit model, may be more appropriate than a continuous linear model.

In a tobit model, the observed dependent variable, y , is equal to the latent variable, y^* , when y^* is within the upper (τ_u) and lower (τ_l) censoring limits, and is otherwise censored at the upper or lower limit when the latent difference is at or outside of those limits (Long 1997):

$$y = \begin{cases} \tau_u, & y^* \geq \tau_u \\ y^* = \alpha + \beta x, & \tau_l < y^* < \tau_u \\ \tau_l, & y^* \leq \tau_l \end{cases} \quad (9.4)$$

Where βx is a vector of attribute coefficients and levels. The regression coefficient, β , represents the marginal change in the latent outcome y^* given a 1-unit change in the level of x .

These assumptions were tested in a series of econometric specification tests (Baltagi 2008; Croissant & Millo 2008). The poolability of the data was tested using Chow's F-test, and the presence of unobserved effects was tested using Wooldridge's test of unobserved effects. A fixed or random effects specification was defined on the basis of Hausman's test, while the presence of specific individual, time or two-way effects was tested with Honda's Lagrange multiplier test. Finally, as the tobit model is based on assumptions of normally distributed and homoscedastic errors (Long 1997), the behaviour of the error term was tested using the Breusch-Pagan test of homoskedasticity and the Anderson-Darling test of normality.

Linear models are also amenable to a latent class modelling approach, where their interpretation as a non-parametric representation of unobserved

heterogeneity is the same as in the DCE analysis. For example, a linear pooled or unobserved effects latent class model can be defined as:

$$y_{it|c} = [\alpha_k + \beta_k x_{it} + e_{it}] \cdot \Pr(c)_i \quad (9.5)$$

Where $y_{it|c}$ is the expected value for individual i at time t , conditional on membership in class c , $[\alpha_k + \beta_k x_{it} + e_{it}]$ is a linear model, and $\Pr(c)_i$ is the probability of class membership as given previously in equation 8.4 (Magidson & Vermunt 2004). A latent class approach can also be extended to a single-bounded tobit model (Brown et al. 2010), but it is currently incompatible with a the double-bounded tobit model.

The dependent variable in each of the different models was the difference in the budget allocation between program A and B, and the parameters were based on the relative differences in attribute levels. The simplest value function was an additive linear main effects differences specification of the form $v = \alpha + \beta_1 \Delta LYg_{it} + \beta_2 \Delta Age_{it} + \beta_3 \Delta U0_{it} + \beta_4 \Delta LE0_{it} + \beta_5 \Delta U1_{it} + \beta_6 \Delta nPats_{it}$, where α was the alternative-specific constant associated with alternative B, ΔLYg was the difference in individual life years gained with treatment between the two alternatives presented to individual i in task t , ΔAge was the difference in age, $\Delta U0$ was the difference in initial utility, $\Delta LE0$ was the difference in life expectancy without treatment, $\Delta U1$ was the difference in utility with/after treatment, and $\Delta nPats$ was the difference in total number patients that could be treated if 100 percent of the budget was allocated to that alternative. All differences were calculated as the level in alternative B less the level in alternative A, and the age and number of patients treated parameters were divided by 10 and 1000, respectively, to re-scale them to a magnitude more comparable with the other parameters in order to improve the chances of model convergence (Long 1997).

An interaction term, interacting the differences in initial and final utility between the two alternative patient groups, was defined as $(1+\Delta U0)(1-\Delta U1)$ and was included in more complex versions of the value function to account for relative differences in quality gain. As the relationship between these terms is not immediately intuitive, the range of possible parameter values for this interaction term is shown Table 9.1.

Table 9.1: Initial and final health state differences interaction values

| | $\Delta U1^{B-A}$ | -0.8 | -0.4 | 0 | 0.4 | 0.8 |
|-------------------|---------------------|------|------|------|------|------|
| | $1-\Delta U1^{B-}$ | | | | | |
| | A | 0.2 | 0.6 | 1.0 | 1.4 | 1.8 |
| $\Delta U0^{B-A}$ | $1+\Delta U0^{B-A}$ | | | | | |
| -0.8 | 1.8 | 0.36 | 1.08 | 1.80 | 2.52 | 3.24 |
| -0.4 | 1.4 | 0.28 | 0.84 | 1.40 | 1.96 | 2.52 |
| 0 | 1.0 | 0.20 | 0.60 | 1.00 | 1.40 | 1.80 |
| 0.4 | 0.6 | 0.12 | 0.36 | 0.60 | 0.84 | 1.08 |
| 0.8 | 0.2 | 0.04 | 0.12 | 0.20 | 0.28 | 0.36 |

The value of the interaction term is maximised when moving from an initial health state where there is a large *negative* difference for patient group B relative to group A (e.g. $U0^A=0.9$, $U0^B=0.1$; $\Delta U0^{B-A}=-0.8$), to a final health state where there is a large *positive* difference between the two patient groups (e.g. $U1^A=0.1$, $U1^B=0.9$; $\Delta U1^{B-A}=0.8$). In other words, situations where patient group B moves from a much worse initial health state to a much better final health state relative to patient group A, and thus gains relatively more quality. The multiplicative interaction avoids collinearity with the main effects, while the $(1+\Delta U0)$ and $(1-\Delta U1)$ terms ensure that relatively more weight is given to the worst and best initial and final health states, respectively. An alternative value function interacted main effects differences with differences in life year gains, weighting the utility associated with differences in specific attributes by the difference in life year gains, consistent with the approach taken by Norman et al. (2013).

The analysis adopted a broadly inclusive significance threshold of 0.10 and parameter p-values were not adjusted for multiple comparisons. Robust standard errors for coefficient estimates were calculated using the ‘sandwich estimator’ (Freedman 2006). The econometric specification tests were conducted with R 2.15.3 using the plm, lmtest and nortest packages, and the models were estimated using LIMDEP 9.0/NLOGIT 4.0.

9.2 Estimating welfare effects

As in the DCE analysis, welfare effects were estimated in terms of compensating variation in the context of a ‘state of the world’ model (Small & Rosen 1981; Ryan 2004; Silva 2004):

$$CV_{a:\Delta LYg} = \frac{1}{\beta_{\Delta LYg}} [\Delta v^0 - \Delta v^1] \quad (9.6)$$

The numeraire, $\beta_{\Delta LYg}$, was the marginal utility of an additional individual life year gained relative to a comparator, and Δv^0 and Δv^1 represented the net difference in utility before and after a change in one or more attribute levels, respectively. Consistent with the interpretation of the CSPC attributes and budget allocations as relative to some comparator, the net utilities are also calculated relative to an implicit comparator. This implicit comparator, though, can be assumed to be identical to the initial state of the scenario under consideration – that is, before any changes in attribute levels – without affecting the interpretation. In this case, the difference in utility between the initial scenario and its implicit comparator (Δv^0) can be assumed to be zero, and Δv^1 represents the net change in utility relative to that original state. A negative CV implies a move to a *more* preferred level (a positive welfare effect), and a positive CV implies a move to a *less* preferred level (a negative welfare effect).

9.2.1 Public vs. agent preferences

A secondary objective of the analysis was to test for heterogeneity between the preferences of self-identified agents and those of the general public. Differences between general public and agent preferences were estimated based on the same model used in the overall analysis, but the value function included an interaction between each parameter and a flag indicating whether or not the respondent self-identified as an agent. If the interactions between specific attributes and agent status were found to be significant, the difference in compensating variation between the general population and agents would be calculated and taken as significant if the 95 percent confidence interval around the difference in CV between agents and the general public did not cross zero.

9.3 Scenario rankings

The CSPC scenarios were ranked by their expected net utility to provide a more holistic sense of the relative attractiveness of each scenario, allowing different attribute levels to vary simultaneously. The predicted net utility of each choice scenario was calculated by weighting the differences in attribute levels between a particular scenario and a reference scenario with all attributes at their middle level, by the coefficients from the identified regression model. Note that this reference scenario was the same one that was used to calculate relative choice probabilities in the DCE scenario rankings, and that it was not actually presented to respondents. Positive relative utility would indicate a scenario was more preferred than the reference scenario, while negative relative utility would indicate a scenario was less preferred than the reference scenario. Unlike the DCE analysis, choice probabilities were not calculated as linear models are not consistent with the estimation of choice probabilities.

Spearman's rho was also calculated to provide a sense of the strength and direction of association between each choice scenario's attribute differences and its relative ranking. The scenarios were ranked by descending utility, so a negative correlation coefficient implies that the relative rank of a scenario improved as an attribute level or difference increased, while a positive correlation coefficient implies that relative rank worsened as an attribute level or difference increased. The CSPC rank and the DCE rank for the same scenario were also compared on the basis of Spearman's rho.

9.4 Distributional preferences

Given the attributes included in each CSPC task, respondents could express a preference for maximising or equalising resources (budget allocations), access (number of patients treated) or outcomes (aggregate QALYs gained). Respondents were classified as strict maximisers if they allocated 100 percent of the budget to one program or the other in each of their choice tasks, and strict equalisers if they chose a 50-50 budget allocation in each of their choice tasks. Respondents could also be classified as equalisers if they chose to equalise the number of patients treated or aggregated QALYs gained in each choice task. As

respondents could only allocate the budget by iterations of 1 percent it was usually not possible to equalise precisely patients or QALYs, so these attributes were taken as equalised if the budget was within 2 percent of the allocation that would precisely equalise patients treated or QALYs gained.¹⁵

In order to distinguish between respondents who saw no difference in the latent utility and were indifferent between alternatives from those who had a strong preference for equality, respondents were only classified as strict equalisers if they also rated distributional concerns as the most important factor in their decisions. The mean number of tasks equalised or maximised by agents and the general public were also compared using t-tests, as were the mean number of responses that equalised the number of patients treated or the number of QALYs gained. The t-tests were conducted using R 2.15.3 (R Core Team 2013).

9.5 Comparison of DCE and CSPC welfare estimates

The CSPC allocations were modelled using a linear model, but the DCE choices were modelled using a non-linear multinomial logit. As such, any observed differences in welfare effects from the two models may reflect, in part, differences in the underlying models assumptions. To test more directly the effect of questionnaire format on the derived estimates of marginal utility and welfare, the CSPC responses were transformed to discrete choices on the basis of which alternative was allocated the majority of the budget and modelled using a multinomial logit comparable to the primary DCE model, including a latent class approach if appropriate. Equal CSPC budget allocations were excluded from the analysis as they did not prioritise either alternative, but this meant that some valid choice data was discarded. Please refer to section 8.2 for more details on the modelling approach used in the DCE analysis.

The equivalence of the attribute coefficients and CV estimates derived from the DCE and CSPC formats for each attribute were tested using a t-test,

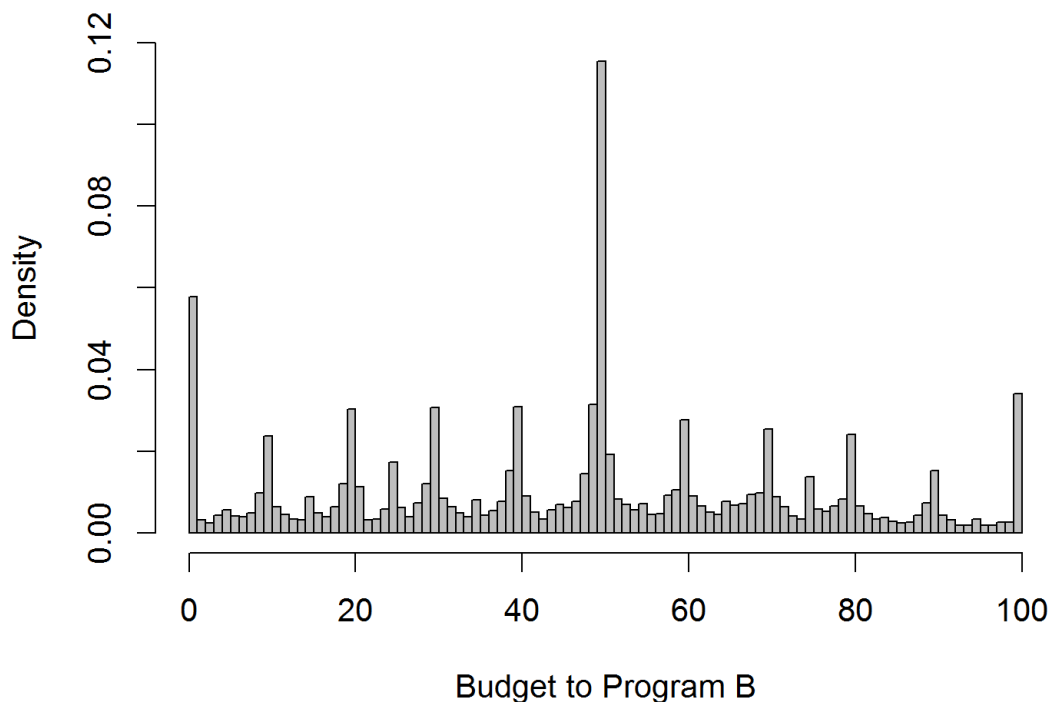
¹⁵ The precise patient- and QALY-equalising budget allocations were calculated as the attribute level in alternative A divided by the sum of the attribute levels in alternative A and B: $nPats^A / (nPats^A + nPats^B)$ and $QALYs^A / (QALYs^A + QALYs^B)$, respectively.

dividing the difference in the estimates from the two choice models by the difference in their standard errors, as per Potoglou (2011). Statistically insignificant results would support the notion of procedural invariance between the two formats, while a significant result may suggest that the elicitation format had a systematic influence on preferences (Carson et al. 1994; Oliver 2013).

9.6 CSPC Results

Chow's F-test of the poolability of the CSPC data rejected the null hypothesis of stable parameters ($p < 0.001$), suggesting an unobserved effects model. This was supported by Wooldridge's test, which rejected the null hypothesis of no unobserved effects ($p < 0.001$). The Hausman test did not reject a random effects specification ($p = 0.99$), while Honda's Lagrange multiplier tests for individual, time and two-way effects rejected the null hypotheses of no significant effects ($p < 0.001$ in all three tests), supporting a two-way random effects specification. Finally, a histogram of the CSPC budget allocations,

Figure 9.1: Primary CSPC budget allocations



shown in Figure 9.1, appeared to confirm censoring in the dependent variable, given distinct clusters at the upper (100% of the budget to Program B) and lower (0% of the budget to Program B) bounds of the budget allocations. In light of this possible clustering, as well as the interpretation of the budget differences as a bounded representation of the differences in latent utility, a tobit was felt to be a theoretically appropriate statistical model for the CSPC responses.

The Breusch-Pagan test did not reject the null hypothesis of homoscedasticity in the tobit error terms ($p=0.18$), but the Anderson-Darling test did reject the assumption of normally distributed errors ($p<0.001$). On this basis,

Figure 9.2: Pooled tobit residuals

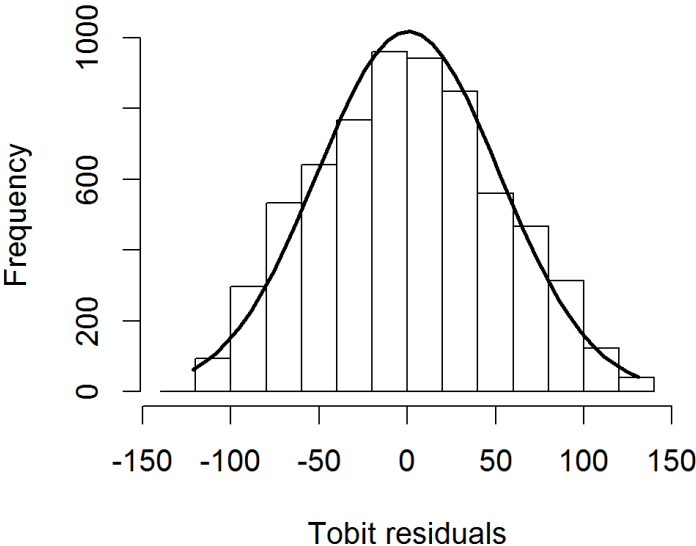


Figure shows a histogram of the pooled tobit residuals with an overlaid normal distribution.

as well as the specification tests suggesting significant individual and time effects, one and two-way random effects linear models were specified. However, as shown in Figure 9.2, the tobit residuals appeared very close to normally distributed, so a tobit specification was also tested. Comparisons of the

goodness of fit of these different models and value function specifications are presented in Appendix 9.1, ordered by improving information criteria.

The initial model was a one-way random effects linear model with continuous main effects differences, analogous to the value function specified for the experimental design. Despite significant two-way effects in the econometric specification tests, a two-way random effects linear model did not find a positive time effect and did not converge, and was excluded from further comparisons. Interacting the difference in individual life years gained with the other main effects improved the fit of the linear model by AICc and BIC, while a

parsimonious version of this model, excluding parameters that were insignificant at a 0.10 threshold, further improved fit by BIC. A latent class linear model based on full and parsimonious versions of this value function did not converge, nor did a random effects double-bounded tobit. A pooled double-bounded tobit with main effects and life year gain interactions improved model fit over the parsimonious linear model, and a parsimonious version of this value function offered the best fit by log-likelihood, AICc and BIC. The coefficients from this parsimonious double-bounded tobit model are shown in Appendix 9.2.

The alternative specific constant in the model was statistically significant, implying a preference for alternative B (the right-hand side of the choice task) independent of attribute levels. The relatively large size of this constant may reflect confounding with the initial and final health state interaction term, as when these differences are zero, the coefficient on the interaction term will be perfectly confounded with the constant (i.e. $\beta_0 + \beta_1[1-\Delta U_0] \times [1+\Delta U_1] = \beta_0 + \beta_1[1] \times [1] = \beta_0 + \beta_1$).

The other attribute coefficients represented the marginal change in the utility of alternative B relative to alternative A, given a marginal change in the continuous attribute level. A positive coefficient indicated that the utility of alternative B increased relative to alternative A as the difference in the level of attribute x in alternative B increased relative to its level in alternative A (i.e. respondents preferred a *higher* level of x). A negative coefficient indicated that relative utility decreased as the relative difference in x increased (i.e. respondents preferred a *lower* level of x). Note that the very large coefficients on the difference in initial and final health states reflect in part the 0-1 scale of those parameters and represent the marginal utility associated with, in effect, the difference in moving from a state equivalent to dead (0.0) to perfect health (1.0). The large and negative coefficient on the initial and final health states interaction term, the value of which increases with a relative gain in quality, would also tend to offset the marginal impact of final health state. The coefficient on the number of patients treated was not significant, but the continuous interaction between life years gained and the number of patients treated was significant and negative, suggesting diminishing returns to aggregate life years gained.

9.6.1 Compensating variations

Attributes with statistically significant compensating variations are detailed in Table 9.2 and illustrated in Figure 9.3. As in the DCE analysis, a negative CV indicated a positive welfare effect, and a positive CV indicated a negative welfare effect. CVs were estimated for an upward change in attribute level relative to a fixed comparator, holding all other differences at zero, but given the linear specification of the value function, CV is necessarily the same for an upward or a downward change; the sign simply reverses for a downward change. This is in contrast to the non-linear, dummy-coded DCE parameters where CV was potentially different for an upward or downward change from the baseline attribute levels.

Table 9.2: CSPC compensating variations by attribute differences

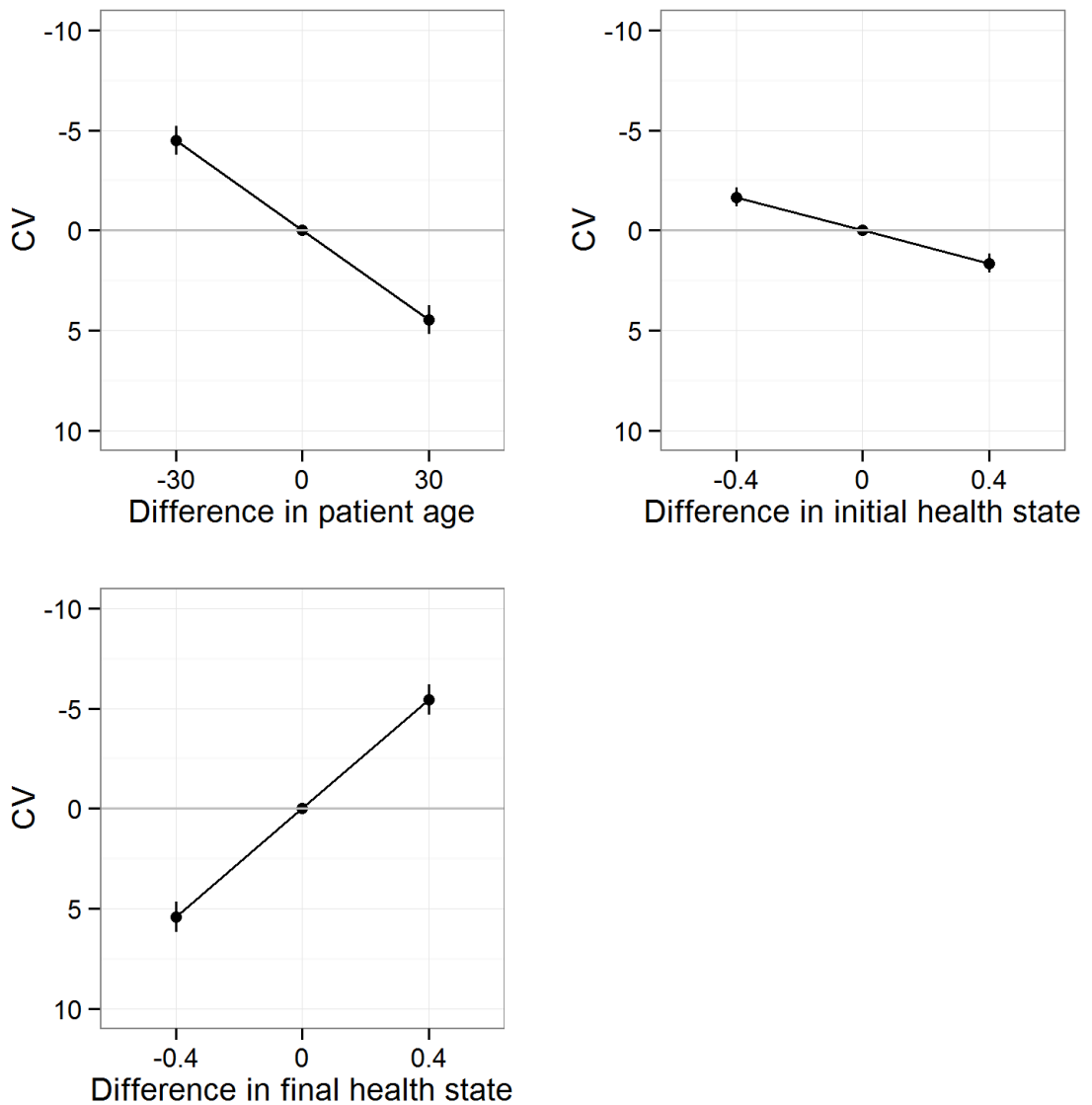
| Attribute difference | CV | Lower 95% CI | Upper 95% CI | Adj. p-value | Sig |
|-------------------------------|-------|--------------|--------------|--------------|-----|
| CV, patient age +30 | 4.49 | 3.77 | 5.20 | <0.001 | *** |
| CV, initial health state +0.4 | 1.65 | 1.17 | 2.13 | <0.001 | *** |
| CV, final health state +0.4 | -5.43 | -6.19 | -4.67 | <0.001 | *** |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

CV=compensating variation; 95% CI = 95% confidence interval. CVs are shown for an increase in the difference between attributes, setting all other attributes differences to zero.

The results suggested that there were positive welfare effects associated with prioritising patient groups that would finish treatment in better final health states, and negative welfare effects associated with prioritising older patient groups or those in better initial health states. There was no statistically significant effect associated with initial life expectancy or, notably, the potential number of patients treated. This was in contrast to the pilot survey where CSPC respondents gave substantial weight to the number of patients treated.

Figure 9.3: CSPC compensating variations by attribute



9.6.2 Marginal effects

As mentioned above, the tobit coefficients were assumed to represent the marginal change in latent utility given a 1-unit change in the difference between attribute levels, consistent with a random utility theory of choice. However, as Long (1997) notes, some researchers are uncomfortable with regarding observed data as a manifestation of a latent process. To directly relate the effect of a change in x to the observed difference in the budget allocation y – not the latent outcome y^* – it is necessary to weight the marginal utility, βx , by the probability of y being censored for a given level x (Long 1997). The marginal effects shown

in Table 9.3 represent the expected change in the *observed* budget difference given a 1-unit change in the in the difference in attribute levels, calculated at the mean attribute differences.

Table 9.3: CSPC double-bounded tobit marginal effects

| Attribute | Marginal effect | Std err | ME/Std err | p-value | Sig | Mean(Δx) |
|-------------------------------------|-----------------|---------|------------|---------|-----|--------------------|
| Constant | 22.65 | 5.27 | 4.30 | <0.001 | *** | |
| Δ Life years gained | 2.92 | 0.19 | 15.76 | <0.001 | *** | 0.48 |
| Δ Patient age / 10 | -4.37 | 0.32 | -13.45 | <0.001 | *** | 2.10 |
| Δ Initial health state | -37.14 | 6.88 | -5.40 | <0.001 | *** | 0.12 |
| Δ Final health state | 64.74 | 7.45 | 8.69 | <0.001 | *** | 0.14 |
| Δ Age: Δ LYg | -0.33 | 0.04 | -7.61 | <0.001 | *** | 4.13 |
| $\Delta U1$: Δ LYg | 2.99 | 0.51 | 5.90 | <0.001 | *** | -0.21 |
| Δn Pats: Δ LYg | -0.35 | 0.04 | -8.51 | <0.001 | *** | -0.49 |
| (1- $\Delta U0$):(1+ $\Delta U1$) | -25.10 | 5.61 | -4.47 | <0.001 | *** | 0.95 |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+'

These marginal effects are very similar to the coefficients reported in Appendix 9.2, but as the interpretation of marginal effects is limited to relating the change in the observed budget allocation given a change in attribute differences, they are largely inconsistent with an understanding of compensating variation as an equalising change in utility. As such, compensating variations based on these marginal effects were not calculated.

9.6.3 Public vs. agent preferences

Differences between general public and agent preferences were estimated using a pooled double-bounded tobit model with clustering, as in the combined analysis. Alternative value function specifications, ranked by information criteria, are shown in Appendix 9.3. The results of the specification with the best fit by AICc and BIC are shown in Appendix 9.4.

All of the attribute-agent interactions were insignificant, even at a 0.10 threshold, suggesting no substantial difference between agent and public preferences. The interaction between the difference in age and agent status was significant at a 0.10 threshold in less parsimonious versions of the value function,

but this interaction was not significant in the final value function. However, given its significance in less parsimonious versions of the value function, the compensating variation and 95 percent confidence intervals associated with a change in age were estimated for agents and the general public. Both groups had positive and statistically significant CVs (negative welfare effects) associated with a 30-year increase in the difference in age: CV was 4.58 (95% CI: 3.85, 5.32) for the general public and 5.71 (95% CI: 4.04, 7.39) for agents. This was suggestive of a slightly stronger preference among agents than the general public for treating younger patients, but the difference was not significant at a 0.05 threshold ($\Delta CV = 1.13$; 95% CI: -0.22, 2.48).

9.6.4 Scenario rankings

The utility of each choice scenario, relative to a hypothetical scenario with all attributes at their middle level, is shown in Table 9.4. This table presents the same 38 choice scenarios that were presented in the DCE ranking of scenarios in Table 8.4, but rather than absolute attribute levels as in the DCE analysis, each scenario here is presented in terms of attribute differences relative to the reference scenario. A negative difference indicates that the attribute level in a particular scenario was *lower* than in the reference scenario, while a positive difference indicates that the attribute level was *higher* than in the reference scenario.

The correlation between each attribute and the overall rank of the scenario is also shown. The scenarios were ranked by descending relative utility, so a negative correlation coefficient indicates that the relative rank of a scenario improved if the difference in an attribute was positive, while a positive correlation coefficient indicates that relative rank worsened if the difference was positive. For comparison purposes, the DCE rank for the same scenario is also shown, along with the correlation between CSPC and DCE rank.

Table 9.4: CSPC scenario rankings by predicted difference in utility

| CSPC Rank | DCE Rank | Δ Age | Δ U0 | Δ LE | Δ U1 | Δ LYg | Δ nPats | Δ Ind QALYs | Δ Agg QALYs | Δ Utility |
|--------------|-------------|--------------|--------------|-------------|--------------|--------------|----------------|--------------------|--------------------|------------------|
| 1 | 7 | -30 | 0 | 0 | 0.4 | 5 | -2400 | 8.50 | -5,150 | 60.73 |
| 2 | 5 | -30 | -0.4 | -4.917 | 0 | 5 | -2400 | 2.53 | -5,747 | 42.36 |
| 3 | 2 | 0 | -0.4 | -4.917 | 0.4 | 5 | 0 | 6.57 | 16,416 | 37.54 |
| 3 | 2 | 0 | -0.4 | -4.917 | 0.4 | 5 | 0 | 6.57 | 16,416 | 37.54 |
| 5 | 1 | -30 | 0 | 0 | 0 | 5 | 2500 | 2.50 | 18,750 | 27.91 |
| 6 | 23 | -30 | 0.4 | 0 | 0.4 | 0 | -2400 | 2.00 | -5,800 | 27.74 |
| 6 | 25 | -30 | 0.4 | 0 | 0.4 | 0 | 0 | 2.00 | 5,000 | 27.74 |
| 8 | 4 | -30 | -0.4 | 5 | 0 | 0 | -2400 | 4.00 | -5,600 | 16.69 |
| 9 | 15 | 0 | 0 | 5 | 0.4 | 0 | -2400 | 6.00 | -5,400 | 14.47 |
| 10 | 10 | 0 | 0 | 5 | 0 | 5 | 0 | 2.50 | 6,250 | 13.12 |
| 11 | 8 | 30 | 0 | 0 | 0.4 | 5 | 2500 | 8.50 | 48,750 | 12.44 |
| 12 | 29 | 30 | 0.4 | 5 | 0.4 | 5 | 2500 | 6.50 | 38,750 | 11.58 |
| 13 | 18 | -30 | 0.4 | -4.917 | 0.4 | -4 | 2500 | -1.60 | -1,750 | 9.44 |
| 14 | 20 | -30 | 0 | -4.917 | -0.4 | 5 | 0 | -1.53 | -3,833 | 9.07 |
| 15 | 11 | -30 | 0 | -4.917 | -0.4 | 5 | 2500 | -1.53 | -1,416 | 4.35 |
| 16 | 22 | -30 | -0.4 | 5 | -0.4 | 0 | 0 | -2.00 | -5,000 | 3.92 |
| 17 | 19 | 0 | 0.4 | -4.917 | 0 | 5 | 2500 | 2.47 | 18,584 | 3.20 |
| 18 | 28 | 0 | -0.4 | -4.917 | 0 | 0 | 0 | 0.03 | 83 | 2.56 |
| 19 | 12 | 30 | -0.4 | -4.917 | 0.4 | 0 | 2500 | 2.07 | 16,582 | 1.19 |
| Ref | Ref | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -2.64 |
| 20 | 9 | 0 | 0 | 5 | 0 | 0 | 2500 | 0.00 | 6,250 | -2.64 |
| 21 | 26 | 0 | 0 | 0 | 0.4 | -4 | 0 | 0.40 | 1,000 | -3.31 |
| 22 | 6 | -30 | 0 | 5 | 0 | -4 | 0 | -2.00 | -5,000 | -5.41 |
| 23 | 13 | 0 | -0.4 | 5 | -0.4 | 5 | 2500 | -1.50 | -1,250 | -5.64 |
| 24 | 16 | 0 | -0.4 | 0 | 0 | -4 | 2500 | 0.00 | 6,250 | -6.27 |
| 25 | 32 | 0 | 0.4 | 5 | 0.4 | -4 | -2400 | -1.60 | -6,160 | -7.80 |
| 26 | 30 | -30 | -0.4 | 5 | -0.4 | -4 | 0 | -2.40 | -6,000 | -7.83 |
| 27 | 23 | 30 | -0.4 | 0 | 0 | 0 | 0 | 2.00 | 5,000 | -11.58 |
| 28 | 21 | 30 | 0.4 | 5 | 0 | 5 | 0 | -1.50 | -3,750 | -11.60 |
| 29 | 14 | -30 | 0 | -4.917 | -0.4 | -4 | 2500 | -2.43 | -5,916 | -13.58 |
| 30 | 37 | 0 | -0.4 | -4.917 | 0 | -4 | -2400 | -1.97 | -6,197 | -13.68 |
| 31 | 17 | 30 | -0.4 | 0 | -0.4 | 5 | -2400 | -1.50 | -6,150 | -15.90 |
| 32 | 35 | 30 | 0 | 5 | 0.4 | -4 | -2400 | 2.40 | -5,760 | -16.76 |
| 33 | 34 | 30 | 0 | -4.917 | 0 | 0 | -2400 | 0.00 | -6,000 | -16.78 |
| 34 | 33 | 0 | 0 | -4.917 | -0.4 | 0 | -2400 | -2.03 | -6,203 | -19.75 |
| 35 | 36 | 0 | 0.4 | -4.917 | 0 | -4 | 0 | -2.03 | -5,083 | -20.44 |
| 36 | 38 | 0 | -0.4 | 0 | -0.4 | -4 | -2400 | -2.40 | -6,240 | -21.29 |
| 37 | 31 | 30 | -0.4 | 5 | -0.4 | 0 | 2500 | -2.00 | -3,750 | -24.36 |
| 38 | 27 | 0 | 0.4 | -4.917 | -0.4 | 0 | 0 | -2.07 | -5,166 | -29.28 |
| Corr. | 0.71 | 0.46 | -0.02 | 0.01 | -0.53 | -0.52 | -0.07 | -0.75 | -0.48 | |

Difference in predicted utility between each choice scenario and a hypothetical reference scenario with all attributes set to their middle level. DCE rank=rank of same scenario in DCE scenario ranking; Δ Age=difference in patient age; Δ U0=difference in initial health state; Δ LE=difference in initial life expectancy; Δ U1=difference in final health state; Δ nPats=difference in potential patients treated; Δ LYg=difference in individual life years gained; Δ Ind QALYs=difference in QALYs gained per patient; Δ Agg QALYs=Difference in aggregate QALYs by group. Corr=Correlation between CSPC rank and attribute difference. The reference scenario is shown in **bold**. The negative relative utility of the reference scenario reflects the significant alternative-specific constant.

The single strongest rank-attribute correlation was with individual QALY gains ($\rho=-0.75$), where the strong negative correlation indicated that as individual QALYs gained increased relative to the reference scenario, so did the rank of that scenario. Each of the top 10 scenarios had positive and relatively large individual QALY gains, while 9 of the bottom 10 scenarios had zero or negative individual QALY gains. Interestingly though, 5 of the top 10 scenarios had negative *aggregate* QALY differences, as larger individual QALY gains accrued to fewer patients than in the reference scenario, and the two scenarios with the largest relative aggregate QALY gains were ranked outside of the top 10. Final health state was the attribute next most strongly correlated with rank, followed by individual life year gains. The number of patients treated attribute had only a very weak association with relative rank. Together these correlations suggested that CSPC respondents emphasised individual over aggregate gains. As in the DCE rankings, age also had a relatively strong association with rank. None of the top 10 scenarios prioritised patients older than the reference scenario, while the patients in 4 of the bottom 10 scenarios – and 6 of the bottom 12 – were older than those in the reference scenario. Severity, either in terms of initial health state or initial life expectancy, had no meaningful impact on the rankings.

Overall, the relative rankings were closely correlated with those from the DCE ($\rho=0.71$), although there were a few notable disagreements. Most apparently, the two scenarios tied at CSPC rank 6 were ranked much less favourably in the DCE. These scenarios involved patients in a relatively good initial health state, and the discordance appeared to stem from an indifference to initial health state in the CSPC and a relatively strong aversion to patients in the best initial health state in the DCE. Similarly, the CSPC had much stronger associations between relative rank and final health state and individual QALY gains than the DCE.

Given the strong association between age and the relative rankings, the scenarios are re-presented in Table 9.5 controlling for the relative difference in age. These results show even more clearly the association between rank and individual QALY gains. The largest net individual QALY gains were consistently ranked at the top of each age category, and relative rank declined in

lockstep with individual QALY gains. They also reinforce the importance of age in the relative rankings. Six of the top 10 scenarios were associated with relatively younger patient groups, while none of the scenarios with relatively older patients were among the top 10, including the scenario with the largest net individual *and* aggregate QALY gains (#11 overall), and the scenario with the second largest aggregate QALY gains (#12 overall).

Table 9.5: CSPC scenario rankings controlling for relative age

| Rank within ΔAge | Overall rank | ΔAge | ΔU0 | ΔLE | ΔU1 | ΔLYg | ΔnPats | ΔInd QALYs | ΔAgg QALYs | ΔUtility |
|-------------------|--------------|------|------|--------|------|------|--------|------------|------------|----------|
| ΔAge = -30 | | | | | | | | | | |
| 1 | 1 | -30 | 0 | 0 | 0.4 | 5 | -2400 | 8.50 | -5,150 | 60.73 |
| 2 | 2 | -30 | -0.4 | -4.917 | 0 | 5 | -2400 | 2.53 | -5,747 | 42.36 |
| 3 | 5 | -30 | 0 | 0 | 0 | 5 | 2500 | 2.50 | 18,750 | 27.91 |
| 4 | 6 | -30 | 0.4 | 0 | 0.4 | 0 | -2400 | 2.00 | -5,800 | 27.74 |
| 5 | 6 | -30 | 0.4 | 0 | 0.4 | 0 | 0 | 2.00 | 5,000 | 27.74 |
| 6 | 8 | -30 | -0.4 | 5 | 0 | 0 | -2400 | 4.00 | -5,600 | 16.69 |
| 7 | 13 | -30 | 0.4 | -4.917 | 0.4 | -4 | 2500 | -1.60 | -1,750 | 9.44 |
| 8 | 14 | -30 | 0 | -4.917 | -0.4 | 5 | 0 | -1.53 | -3,833 | 9.07 |
| 9 | 15 | -30 | 0 | -4.917 | -0.4 | 5 | 2500 | -1.53 | -1,416 | 4.35 |
| 10 | 16 | -30 | -0.4 | 5 | -0.4 | 0 | 0 | -2.00 | -5,000 | 3.92 |
| 11 | 23 | -30 | 0 | 5 | 0 | -4 | 0 | -2.00 | -5,000 | -5.41 |
| 12 | 27 | -30 | -0.4 | 5 | -0.4 | -4 | 0 | -2.40 | -6,000 | -7.83 |
| 13 | 30 | -30 | 0 | -4.917 | -0.4 | -4 | 2500 | -2.43 | -5,916 | -13.58 |
| ΔAge = 0 | | | | | | | | | | |
| 1 | 3 | 0 | -0.4 | -4.917 | 0.4 | 5 | 0 | 6.57 | 16,416 | 37.54 |
| 2 | 3 | 0 | -0.4 | -4.917 | 0.4 | 5 | 0 | 6.57 | 16,416 | 37.54 |
| 3 | 9 | 0 | 0 | 5 | 0.4 | 0 | -2400 | 6.00 | -5,400 | 14.47 |
| 4 | 10 | 0 | 0 | 5 | 0 | 5 | 0 | 2.50 | 6,250 | 13.12 |
| 5 | 17 | 0 | 0.4 | -4.917 | 0 | 5 | 2500 | 2.47 | 18,584 | 3.20 |
| 6 | 18 | 0 | -0.4 | -4.917 | 0 | 0 | 0 | 0.03 | 83 | 2.56 |
| <i>Ref</i> | <i>Ref</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0 | -2.64 |
| 7 | 20 | 0 | 0 | 5 | 0 | 0 | 2500 | 0.00 | 6,250 | -2.64 |
| 8 | 22 | 0 | 0 | 0 | 0.4 | -4 | 0 | 0.40 | 1,000 | -3.31 |
| 9 | 24 | 0 | -0.4 | 5 | -0.4 | 5 | 2500 | -1.50 | -1,250 | -5.64 |
| 10 | 25 | 0 | -0.4 | 0 | 0 | -4 | 2500 | 0.00 | 6,250 | -6.27 |
| 11 | 26 | 0 | 0.4 | 5 | 0.4 | -4 | -2400 | -1.60 | -6,160 | -7.80 |
| 12 | 31 | 0 | -0.4 | -4.917 | 0 | -4 | -2400 | -1.97 | -6,197 | -13.68 |
| 13 | 35 | 0 | 0 | -4.917 | -0.4 | 0 | -2400 | -2.03 | -6,203 | -19.75 |
| 14 | 36 | 0 | 0.4 | -4.917 | 0 | -4 | 0 | -2.03 | -5,083 | -20.44 |
| 15 | 37 | 0 | -0.4 | 0 | -0.4 | -4 | -2400 | -2.40 | -6,240 | -21.29 |
| 16 | 39 | 0 | 0.4 | -4.917 | -0.4 | 0 | 0 | -2.07 | -5,166 | -29.28 |
| ΔAge = +30 | | | | | | | | | | |

| | | | | | | | | | | |
|----------|----|----|------|--------|------|----|-------|-------|--------|--------|
| 1 | 11 | 30 | 0 | 0 | 0.4 | 5 | 2500 | 8.50 | 48,750 | 12.44 |
| 2 | 12 | 30 | 0.4 | 5 | 0.4 | 5 | 2500 | 6.50 | 38,750 | 11.58 |
| 3 | 19 | 30 | -0.4 | -4.917 | 0.4 | 0 | 2500 | 2.07 | 16,582 | 1.19 |
| 4 | 28 | 30 | -0.4 | 0 | 0 | 0 | 0 | 2.00 | 5,000 | -11.58 |
| 5 | 29 | 30 | 0.4 | 5 | 0 | 5 | 0 | -1.50 | -3,750 | -11.60 |
| 6 | 32 | 30 | -0.4 | 0 | -0.4 | 5 | -2400 | -1.50 | -6,150 | -15.90 |
| 7 | 33 | 30 | 0 | 5 | 0.4 | -4 | -2400 | 2.40 | -5,760 | -16.76 |
| 8 | 34 | 30 | 0 | -4.917 | 0 | 0 | -2400 | 0.00 | -6,000 | -16.78 |
| 9 | 38 | 30 | -0.4 | 5 | -0.4 | 0 | 2500 | -2.00 | -3,750 | -24.36 |

Difference in predicted utility between each choice scenario and a hypothetical reference scenario with all attributes set to their middle level, grouped by relative difference in age. DCE rank=rank of same scenario in DCE analysis; Δ Age=difference in patient age; Δ U0=difference in initial health state; Δ LE=difference in initial life expectancy; Δ U1=difference in final health state; Δ nPats=difference in potential patients treated; Δ LYg=difference in individual life years gained; Δ Ind QALYs=difference in QALYs gained per patient; Δ Agg QALYs=Difference in aggregate QALYs by group. The reference scenario is shown in **bold**.

The rankings were also stratified by initial life expectancy to explore the relationship with life years gained (not shown). As was observed in the DCE rankings, scenarios with relatively greater individual life year gains were consistently ranked more favourably in scenarios with relatively shorter life expectancies, while there was no clear pattern in scenarios with relatively greater life expectancy. This appeared contrary to Harris' (1985) argument that the duration of benefit should be immaterial to prioritising decisions on the grounds that an individual with a short life expectancy may place the same value on their remaining time as an individual with a much longer life expectancy. If Harris' argument held, one would expect the observed pattern to be reversed, with respondents indifferent to the duration of benefit in patients with the shortest initial life expectancy. This somewhat counter-intuitive result, though, may reflect diminishing returns to life year gains in patients with relatively longer initial life expectancies.

9.6.5 Distributional preferences

Although the modal CSPC response (12% of all CSPC tasks) was an equal budget allocation to each program (zero budget difference), only 8 of 658 CSPC respondents (1.2%, excluding the 4 respondents deemed non-informative), all from the general population sample, equalised budgets between the two alternatives in every choice task. Likewise, even though the second most frequent response (9% of all CSPC tasks) was a maximum budget allocation to one program or the other (+100 or -100 budget difference), no CSPC respondents

maximised budgets in every task. The distribution of respondents by the number of tasks equalised or maximised (excluding the repeated task) is shown in Table 9.6.

Table 9.6: Respondents by the number of tasks equalising or maximising the budget allocation

| Tasks | Respondents equalising budgets | | | Respondents maximising budgets | | |
|----------------------------------|--------------------------------|-------------|--------------|--------------------------------|-------------|--------------|
| | Agents | Public | Combined | Agents | Public | Combined |
| 0 | 30 (52.6%) | 288 (47.9%) | 318 (48.3%) | 49 (86.0%) | 475 (79.0%) | 524 (79.6%) |
| 1 | 14 (24.6%) | 153 (25.5%) | 167 (25.4%) | 4 (7.0%) | 80 (13.3%) | 84 (12.8%) |
| 2 | 9 (15.8%) | 78 (13.0%) | 87 (13.2%) | 2 (3.5%) | 28 (4.7%) | 30 (4.6%) |
| 3 | 1 (1.8%) | 31 (5.2%) | 32 (4.9%) | 1 (1.8%) | 8 (1.3%) | 9 (1.4%) |
| 4 | 3 (5.3%) | 20 (3.3%) | 23 (3.5%) | 1 (1.8%) | 3 (0.5%) | 4 (0.6%) |
| 5 | 0 (0.0%) | 9 (1.5%) | 9 (1.4%) | 0 (0.0%) | 5 (0.8%) | 5 (0.8%) |
| 6 | 0 (0.0%) | 4 (0.7%) | 4 (0.6%) | 0 (0.0%) | 2 (0.3%) | 2 (0.3%) |
| 7 | 0 (0.0%) | 3 (0.5%) | 3 (0.5%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| 8 | 0 (0.0%) | 3 (0.5%) | 3 (0.5%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| 9 | 0 (0.0%) | 4 (0.7%) | 4 (0.6%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| 10 | 0 (0.0%) | 8 (1.3%) | 8 (1.2%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| Total | 57 (8.7%) | 601 (91.3%) | 658 (100.0%) | 57 (8.7%) | 601 (91.3%) | 658 (100.0%) |
| Mean tasks equal. or max. | 0.82 | 1.18 | 1.15 | 0.26 | 0.35 | 0.34 |
| p-value | 0.03 | | | 0.43 | | |

Excludes 4 general public respondents who did not move the slider in any of their choices and finished the questionnaire in less than one-half the median completion time.

The table shows that neither maximising nor equalising strategies were particularly common among respondents: 48 percent of respondents did not equalise budgets in any tasks, and 80 percent of respondents did not maximise budgets in any tasks. Overall, the average respondent equalised 1.15 tasks and maximised 0.34 tasks. Agents had a significantly lower mean number of tasks equalised than the general population sample, but there was no significant difference in the number of tasks maximised. Respondents who somewhat or strongly disagreed with an inevitable need for rationing in healthcare appeared to equalise slightly but significantly more tasks than respondents who somewhat or strongly agreed (1.48 vs. 1.07, $p=0.04$), but there was no significant difference in

the number of tasks maximised ($p=0.58$). Only 2 of the 8 strict equalising respondents also rated distributional concerns as the most important factor in their decisions, but 4 of the 8 equalisers had a completion time less than half of the median CSPC completion time. It is not clear, though, whether an egalitarian preference may have simplified the CSPC tasks and led to faster completion times, or whether equalisation may have been adopted as a simplifying heuristic to complete the tasks more quickly.

The equalising and maximising behaviours observed here were substantially different than observed by Schwappach (2003) in a similar CSPC in a German setting. He reported that 11 percent of all respondents equalised budget allocations in every task, compared to 1 percent here, and only 3 percent of all allocations maximised the budget to one program or the other, compared to 9 percent here. As in the pilot survey, these low rates suggest that respondents were not using equal budget allocations to avoid making difficult allocation choices.

In addition to equalising budgets, CSPC respondents could also allocate the budget in each task so as to equalise the number of patients treated or the aggregate QALYs gained between the two alternatives. The distribution of respondents by the number of tasks where patients or QALYs were equalised (excluding the repeated task) is shown in Table 9.7.

Table 9.7: Respondents by the number of tasks equalising patients treated or QALYs gained

| Tasks | Respondents equalising patients treated | | | Respondents equalising QALYs gained | | |
|-------|---|-------------|-------------|-------------------------------------|-------------|-------------|
| | Agents | Public | Combined | Agents | Public | Combined |
| 0 | 41 (71.9%) | 350 (58.2%) | 391 (59.4%) | 41 (71.9%) | 440 (73.2%) | 481 (73.1%) |
| 1 | 12 (21.1%) | 185 (30.8%) | 197 (29.9%) | 11 (19.3%) | 130 (21.6%) | 141 (21.4%) |
| 2 | 3 (5.3%) | 48 (8.0%) | 51 (7.8%) | 4 (7.0%) | 27 (4.5%) | 31 (4.7%) |
| 3 | 1 (1.8%) | 15 (2.5%) | 16 (2.4%) | 0 (0.0%) | 2 (0.3%) | 2 (0.3%) |
| 4 | 0 (0.0%) | 2 (0.3%) | 2 (0.3%) | 0 (0.0%) | 1 (0.2%) | 1 (0.2%) |
| 5 | 0 (0.0%) | 1 (0.2%) | 1 (0.2%) | 1 (1.8%) | 0 (0.0%) | 1 (0.2%) |
| 6 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| 7 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| 8 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| 9 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 1 (0.2%) | 1 (0.2%) |
| 10 | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |

| | | | | | | |
|--|-------------|-------------|-------------|-------------|-------------|-------------|
| Total | 57 (8.7%) | 601 (91.3%) | 658 (100%) | 57 (8.7%) | 601 (91.3%) | 658 (100%) |
| Mean tasks with equal patients or QALYs | 0.37 | 0.56 | 0.55 | 0.42 | 0.34 | 0.34 |
| Adj. p-value | 0.14 | | 0.39 | | | |

Tasks were classified as equalised if the budget allocation was within 2 percent of the precise equalising allocation. Excludes 4 general public respondents who did not move the slider in any of their choices and finished the questionnaire in less than one-half the median completion time.

Agents appeared slightly less likely than the general population sample to equalise patients in their budget allocations, but slightly more likely equalise QALYs. However, neither difference was statistically significant (adjusted $p=0.14$ and 0.39 , respectively), and the majority of respondents did not equalise patients treated or aggregate QALYs gained in any of their choices. As noted, a task was classified as equalised if the budget was within 2 percent of the precise equalising allocation for either attribute. This margin of error meant that the distribution shown in Table 9.7 may, if anything, slightly overstate the true proportion of equalised tasks.

9.6.6 Comparison of preferences from the DCE and CSPC

The transformed CSPC discrete choices, based on the alternative that was allocated the majority of the budget, were modelled using a latent class multinomial logit model. Twelve percent of all CSPC responses assigned an equal 50-50 budget allocation and were excluded from this analysis as they did not prioritise either alternative.¹⁶ The initial value function was the same as used in the DCE analysis, and included continuous life years gained, dummy-coded main effects, a continuous interaction between initial and final health state, and continuous interactions between attribute main effects and individual life years gained. A 2-class model with the full value function did not converge, but a more parsimonious specification, excluding less plausible interactions between life years gained and initial health state and initial life expectancy did converge. A third specification, excluding insignificant interactions between life years

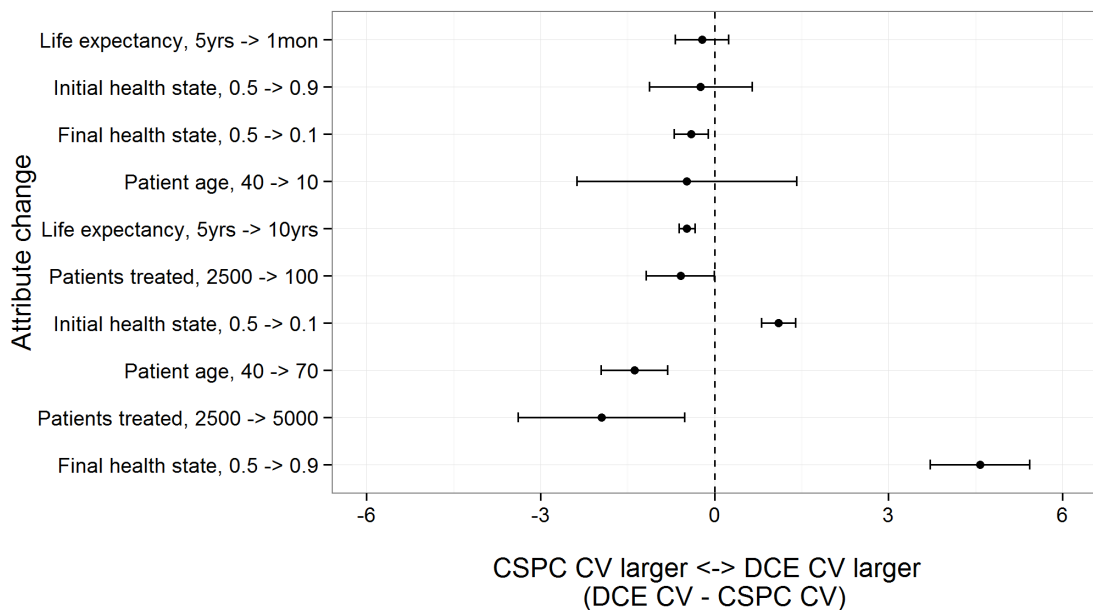
¹⁶ It would also have been possible to randomly choose one of the two alternatives for each equal budget allocation, but it was felt that this may have misrepresented the preferences of respondents who explicitly preferred equality in resources, or who consciously 'chose not to choose.'

gained and age and patients treated further improved model fit by BIC. All the remaining parameters in this specification were significant at a 0.10 threshold in at least one of the two latent classes. A 3-class finite mixture model with the same parsimonious value function specification also converged, but had very large standard errors in one class and was rejected. The results of the parsimonious 2-class model are shown in Appendix 9.5, along with the results from the 2-class DCE latent class model from the primary analysis, previously described in Chapter 8.

T-tests comparing the DCE and CSPC latent class coefficients suggested that the differences between most of the coefficients in the two models were statistically significant at a 0.05 threshold, even after adjusting for multiple comparisons. The most notable exception was the insignificant difference in the coefficients on both dummy-coded age parameters. However, given interaction effects included in both models, it was difficult to predict differences in welfare effects based on differences in the coefficients alone.

The differences in the compensating variations between the two questionnaire formats are illustrated in Figure 9.4, arranged by increasing absolute difference.

Figure 9.4: CV differences by attribute change, DCE vs. CSPC



The figure suggested that there was considerable variation between the two formats, ranging from relatively small and statistically insignificant differences associated with a downward change in life expectancy or an upward change in initial health state, to much larger and statistically significant differences associated with upward changes in the number of patients treated and final health state. The format-specific compensating variations and 95 percent confidence intervals, as well as the marginal utility of an additional individual life year, are shown in Table 9.8.

Table 9.8: Compensating variations and differences by attribute and questionnaire format

| Change | DCE CV (95% CI) | CSPC CV (95% CI) | Difference (95% CI) |
|-------------------------------------|-------------------------|-------------------------|---------------------------------------|
| Patient age, 40 → 10 | -4.36 (-7.45, -1.26) | -3.87 (-5.08, -2.67) | -0.48 (-2.38, 1.41) |
| Patient age, 40 → 70 | 2.91 (0.91, 4.91) | 4.29 (2.87, 5.72) | -1.38 (-1.96, -0.81) |
| Initial health state, 0.5 → 0.1 | -0.57 (-1.63, 0.48) | -1.67 (-2.43, -0.91) | 1.10 (0.81, 1.39) |
| Initial health state, 0.5 → 0.9 | 1.41 (-0.55, 3.36) | 1.65 (0.58, 2.72) | -0.24 (-1.13, 0.65) |
| Life expectancy, 5yrs → 1mon | 3.57 (1.82, 5.32) | 3.79 (2.49, 5.08) | -0.22 (-0.68, 0.24) |
| Life expectancy, 5yrs → 10yrs | -0.77 (-1.3, -0.25) | -0.30 (-0.95, 0.36) | -0.48 (-0.34, -0.61) |
| Final health state, 0.5 → 0.1 | 2.88 (1.34, 4.43) | 3.29 (2.04, 4.54) | -0.41 (-0.7, -0.11) |
| Final health state, 0.5 → 0.9 | 0.71 (-1.27, 2.69) | -3.86 (-4.99, -2.74) | 4.58 (3.72, 5.43) |
| Total patients treated, 2500 → 100 | -0.60 (-2.03, 0.83) | -0.01 (-0.85, 0.84) | -0.59 (-1.18, -0.01) |
| Total patients treated, 2500 → 5000 | -4.2 (-6.55, -1.86) | -2.25 (-3.16, -1.34) | -1.95 (-3.39, -0.52) |
| Marginal utility of 1 LY gained | 0.28 (0.16, 0.40) | 0.12 (0.08, 0.15) | 0.17 (0.08, 0.25) |

95% CI=95% confidence interval. Differences calculated as CSPC CV less DCE CV. A negative CV indicates a positive welfare effect.

Perhaps the single most notable difference was over final health state, where there was no significant welfare effect associated with prioritising patients that would finish treatment in the best final health state over those who would finish in a moderate final health state in the DCE, but a relatively strong and statistically significant positive effect in the CSPC. Likewise, there was no

significant effect in the DCE associated with prioritising patients in the worst initial health state over those in a moderate state, but a significant positive effect in the CSPC. Together these suggest a greater emphasis on quality improvement in the CSPC compared to the DCE. In contrast, there was a significantly stronger welfare effect in the DCE associated with prioritising the largest patient group. This was somewhat counter to expectations, as evidence from the pilot survey suggested the CSPC may have been more likely to emphasise the total number of patients treated as this attribute changed with the budget allocation. Based on the significant difference in the marginal utility of life year gains, respondents to the DCE also appeared to value an additional life year more highly than respondents to the CSPC.

Interestingly, there were negative welfare effects associated with prioritising patients with the shortest initial life expectancy in both questionnaires. This contradicts expectations from the empirical ethics review, and highlights a difference between the tobit and multinomial logit model results. Initial life expectancy was insignificant in the linear tobit analysis, but the non-linear analysis of the transformed CSPC responses appeared to reveal a statistically significant effect. It is important to recognise, however, that the transformed CSPC responses excluded equal budget allocations, so it not clear whether this was a true effect revealed by the non-linear analysis or simply an artefact of the subset of responses included in the analysis.

9.7 Discussion of CSPC results

The results of the primary tobit analysis of CSPC responses suggested that prioritising younger patients, those in worse initial health states, and those that would finish in better final health states, as well as those that would gain more individual life years, were associated with positive welfare effects. There was no statistically significant effect over the range of life expectancy tested or, unexpectedly, over the potential number of patients treated. This was in contrast to the pilot survey, where CSPC respondents gave substantial weight to the number of patients treated. Based on this pilot result, it was hypothesised that CSPC respondents may have taken the fact that the number of patients treated

changed as they moved the budget slider as a cue to focus on this attribute, leading to a prominence effect. The insignificance of the number of patients treated attribute, and the lack of any association between this attribute and the relative ranking of each CSPC scenario, appeared to discount this notion. Larger aggregate QALY gains were associated with more favourable rankings, but this association was not as strong as the association between rank and individual QALY gains, or between rank and individual life year gains.

The proportional relationship between the relative budget and the number of patients treated in any alternative highlighted the fact that giving greater priority to one group meant that the other group must necessarily receive lower priority. However, it may also have complicated the interpretation of this attribute by transforming it from an input parameter to an implicit response variable – relatively more patients were treated in alternatives that respondents preferred. In this sense, allowing the number of patients treated attribute to change may have shifted the CSPC towards a form of a person trade-off (PTO) task. It is not clear how the significance of the number of patients treated attribute might have changed if it had been held fixed. In light of this possible shift in the interpretation of the CSPC, however, it may be informative to contrast the results here with the PTO approach reviewed in Chapter 3. Both the CSPC and the PTO clearly highlight the inter-personal trade-offs inherent in healthcare priority setting, but it was argued that the CSPC may have an advantage in eliciting preferences over these trade-offs in a more intuitive, less discomfoting manner. This appeared to be supported by the results. In contrast to the 91 percent of PTO respondents reported by Damschroder (2007), and the 32 percent reported by Nord (1995a), who refused to make any trade-offs between patient groups and set the person-equivalents equal, only 12 percent of all CSPC tasks equalised the budget allocations between alternatives, and only 1 percent of all respondents equalised the budget allocation in every one of their choices.

When the CSPC responses were transformed from linear differences to discrete choices, a negative and statistically significant welfare effect associated with prioritising patients with the shortest untreated life expectancy emerged, as did a positive and statistically significant effect associated with prioritising the

largest patient groups. However, the strength of the effect around prioritising the largest patient group in the CSPC was still significantly less than the corresponding effect in the DCE. Overall, CSPC respondents appeared to put significantly more weight on initial and final quality of life, while DCE respondents put significantly more weight on maximising the number of patients treated and individual life year gains, controlling for the level of the other attributes. This result appears consistent with the characterisation of the DCE as a more competitive task, as these two attributes were easy to compare between alternatives in order to identify a ‘winning’ alternative. Interestingly, this appears not only to contradict the idea of a prominence effect in the, but suggests that CSPC respondents were in fact less likely than DCE respondents to consider aggregate outcomes in their choices.

The CSPC allowed respondents to express specific distributional preferences, including preferences for maximising resources and/or outcomes, or for the equality of resources (budget), opportunity (patients treated) or outcomes (QALYs gained). As it was, however, only a handful of CSPC respondents expressed preferences for strict resource equality. As was also noted in the discussion of the pilot elicitation in Chapter 5, the low incidence of egalitarian behaviour was surprising, as it was expected that respondents would view in equality in at least one of these aspects as a heuristic for a fair allocation. Furthermore, given the relatively high proportion of fast responders among these respondents, and the relatively low proportion of these respondents who also ranked distributional concerns as the most important factor in their choices, it is likely that at least some of the egalitarian responses were the result of a simplifying heuristic rather than a considered preference for equality. It is not clear, though, whether an equalising decision rule may have led to fast completion times, or whether a desire to complete the questionnaire as quickly as possible may have led to an equalising decision rule.

Agents were slightly but significantly less likely than the general population sample to equalise budget allocations, while respondents who disagreed within an inevitable need for healthcare rationing were significantly more likely to equalise budget allocations. This likely reflects the underlying attitudes of the two groups: agents were more likely to agree with a need for

rationing, and therefore may have been somewhat more willing to prioritise specific patient groups, while those who disagreed were less willing to see one group advantaged relative to another in terms of resources allocated. No respondents maximised the budget allocations to one group or the other in all 10 choice tasks, perhaps reflecting a general aversion to extreme distributions in the allocation of societal resources suggested by Schwappach and Strasmann (2006), and consistent with the fairness principle of ‘everybody gets something and nobody gets nothing’ noted by Giacomini et al. (2012).

The insignificance of the number of patients treated, combined with the relatively low rate of resource-equalising allocations, appeared to imply that CSPC respondents were not necessarily concerned with maximising aggregate gains or equalising resources between groups. Rather, based on the results from the scenario rankings, they appeared more concerned with maximising individual QALY gains, particularly to those patients that would finish treatment in better health states. Although there is nothing irrational or invalid about this preference, it may suggest that respondents may have been reducing the abstract, macro-level allocation problems to more comprehensible two-person analogies (Giacomini et al. 2012; Ryan 2009). This may have been exacerbated by the use of QALY graphs in the elicitation tasks that illustrated the gains at the individual rather than the program level (see Appendix 6.3 for a sample graph). In this light, it is not clear whether the insignificance of the number of patients treated reflected a genuine indifference to aggregate outcomes, or a simplifying approach to the tasks. The implications of these preferences results for healthcare priority setting, as well as for choosing between the CSPC and DCE as a preferred format for eliciting preferences, will be discussed in the next chapter.

Appendix 9.1: Alternative CSPC models and utility functions, by improving information criteria

| Model and utility function | k | LL | AICc | BIC |
|---|----|---------------|--------------|--------------|
| 1.0) 1-way linear random effects; continuous main effects differences $\Delta v = \Delta LYg + \Delta Age + \Delta U0 + \Delta LE + \Delta U1 + \Delta nPats$ | 7 | -34815 | 71105 | 75459 |
| 2.0) 1-way linear random effects; continuous main effects differences + LYg interactions + (1- $\Delta U0$): $\Delta U1$ interaction $\Delta v = \Delta LYg + \Delta Age + \Delta U0 + \Delta LE + \Delta U1 + \Delta nPats + \Delta Age:\Delta LYg + \Delta U0:\Delta LYg + \Delta LE:\Delta LYg + \Delta U1:\Delta LYg + \Delta nPats:\Delta LYg + (1-\Delta U0):(1+\Delta U1)$ | 13 | -34742 | 70974 | 75366 |
| 2.1) Parsimonious 1-way linear random effects; continuous main effects differences + interactions $\Delta v = \Delta LYg + \Delta Age + \Delta U0 + \Delta U1 + \Delta Age:\Delta LYg + \Delta U1:\Delta LYg + \Delta nPats:\Delta LYg + (1-\Delta U0):(1+\Delta U1)$ | 9 | -34747 | 70976 | 75342 |
| 3.0) Pooled double-bounded tobit; log-linear main effects differences $\Delta v = \Delta \log(LYg) + \Delta \log(Age) + \Delta \log(U0) + \Delta \log(LE) + \Delta \log(U1) + \Delta \log(nPats)$ | 7 | -33499 | 67014 | 67068 |
| 3.1) Pooled double-bounded tobit; parsimonious log-linear main effects differences $\Delta v = \Delta \log(LYg) + \Delta \log(Age) + \Delta \log(U0) + \Delta \log(LE) + \Delta \log(U1)$ | 6 | -33499 | 67012 | 67059 |
| 4.0) Pooled double-bounded tobit; main effects differences $\Delta v = \Delta LYg + \Delta Age + \Delta U0 + \Delta LE + \Delta U1 + \Delta nPats$ | 7 | -33418 | 66851 | 66906 |
| 5.0) Pooled double-bounded tobit; main effects differences + LYg interactions + (1- $\Delta U0$): $\Delta U1$ interaction $\Delta v = \Delta LYg + \Delta Age + \Delta U0 + \Delta LE + \Delta U1 + \Delta nPats + \Delta Age:\Delta LYg + \Delta U0:\Delta LYg + \Delta LE:\Delta LYg + \Delta U1:\Delta LYg + \Delta nPats:\Delta LYg + (1-\Delta U0):(1+\Delta U1)$ | 13 | -33357 | 66743 | 66837 |
| 5.1) Pooled double-bounded tobit; parsimonious main effects differences + interactions $\Delta v = \Delta LYg + \Delta Age + \Delta U0 + \Delta U1 + \Delta Age:\Delta LYg + \Delta U0:\Delta LYg + \Delta U1:\Delta LYg + \Delta nPats:\Delta LYg + (1-\Delta U0):(1+\Delta U1)$ | 9 | -33361 | 66741 | 66809 |

k=parameters, including alternative specific constant; LL=Log-likelihood; AICc= Akaike information criterion, with correction for finite sample size; BIC=Bayesian information criterion. Only models and value functions associated with an improvement in LL, AICc or BIC over the previous specification are shown. The minimum overall log-likelihood, AICc and BIC are shown in bold.

Appendix 9.2: CSPC double-bounded tobit coefficients

| Attribute | Coefficient | Std err | Coef of var | β /Std err | p-value | Sig |
|--|-------------|---------|-------------|------------------|---------|-----|
| Constant | 24.44 | 5.69 | 0.233 | 4.29 | <0.001 | *** |
| Δ Life years gained | 3.15 | 0.20 | 0.063 | 15.55 | <0.001 | *** |
| Δ Patient age / 10 | -4.71 | 0.35 | -0.074 | -13.33 | <0.001 | *** |
| Δ Initial health state | -40.07 | 7.44 | -0.186 | -5.39 | <0.001 | *** |
| Δ Final health state | 69.85 | 8.10 | 0.116 | 8.62 | <0.001 | *** |
| Δ Age: Δ LYg | -0.36 | 0.05 | -0.139 | -7.57 | <0.001 | *** |
| Δ U1: Δ LYg | 3.23 | 0.55 | 0.170 | 5.87 | <0.001 | *** |
| Δ nPats: Δ LYg | -0.38 | 0.04 | -0.105 | -8.46 | <0.001 | *** |
| $(1-\Delta$ U0): $(1+\Delta$ U1) | -27.08 | 6.07 | -0.224 | -4.46 | <0.001 | *** |
| Sigma | 55.42 | 0.87 | 0.016 | 63.48 | <0.001 | *** |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10='+' | | | | | | |

Appendix 9.3: Alternative CSPC public-agent interaction value functions, by improving information criteria

| Attribute | p-value, model 1 | p-value, model 2 | p-value, model 3 | p-value, model 4 | p-value, model 5 |
|----------------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Δ Life years gained | 0.140 | 0.000 | 0.000 | 0.000 | 0.000 |
| Δ Patient age / 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Δ Initial health state | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Δ Life expectancy | 0.254 | | | | |
| Δ Final health state | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Δ Patients treated | 0.711 | | | | |
| (1-ΔU0):(1+ΔU1) | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 |
| ΔAge:ΔLYg | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ΔU0:ΔLYg | 0.789 | | | | |
| ΔLE:ΔLYg | 0.831 | | | | |
| ΔU1:ΔLYg | 0.066 | | | | |
| ΔnPats:ΔLYg | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| (1-ΔU0):(1+ΔU1):ΔLYg | 0.861 | | | | |
| ΔLYg:Agent | 0.047 | 0.213 | | | |
| ΔAge:Agent | 0.102 | 0.005 | 0.036 | 0.055 | 0.101 |
| ΔU0:Agent | 0.701 | | | | |
| ΔLE:Agent | 0.2313 | | | | |
| ΔU1:Agent | 0.2496 | 0.2398 | | | |
| ΔnPats:Agent | 0.9892 | | | | |
| (1-ΔU0):(1+ΔU1):Agent | 0.8302 | 0.2733 | | | |
| ΔAge:ΔLYg:Agent | 0.9168 | | | | |
| ΔU0:ΔLYg:Agent | 0.0356 | 0.0332 | 0.088 | 0.1079 | |
| ΔLE:ΔLYg:Agent | 0.1404 | | | | |
| ΔU1:ΔLYg:Agent | 0.1649 | | | | |
| ΔnPats:ΔLYg:Agent | 0.5456 | | | | |
| (1-ΔU0):(1+ΔU1):ΔLYg:Agent | 0.0304 | 0.0736 | 0.2039 | | |
| Constant | 0.0071 | 0.000 | 0.000 | 0.000 | 0.000 |
| LL | -33346 | -33354 | -33356 | -33357 | -33359 |
| AICc | 66747 | 66737 | 66737 | 66737 | 66738 |
| BIC | 66930 | 66839 | 66818 | 66812 | 66806 |

Specifications are based on a double-bounded tobit. LL=Log-likelihood; AICc= Akaike information criterion with correction for finite sample size; BIC=Bayesian information criterion. Only value functions associated with an improvement in LL, AICc or BIC over the previous specification are shown. The overall minimum log-likelihood, AICc and BIC are shown in bold. The p-value of the likelihood ratio $\Pr(\chi^2)$ is shown for relative to the full specification.

Appendix 9.4: CSPC double-bounded tobit with agent interactions

| Attribute | Coefficient | Std err | β /Std err | p-value | Sig |
|---|-------------|---------|------------------|---------|-----|
| Constant | 24.41 | 5.70 | 4.29 | <0.001 | *** |
| Δ Life years gained (Δ LYg) | 3.15 | 0.20 | 15.55 | <0.001 | *** |
| Δ Patient age / 10 | -4.82 | 0.36 | -13.26 | <0.001 | *** |
| Δ Initial health state | -40.04 | 7.44 | -5.38 | <0.001 | *** |
| Δ Final health state | 69.79 | 8.11 | 8.61 | <0.001 | *** |
| (1- Δ U0):(1+ Δ U1) | -27.05 | 6.07 | -4.46 | <0.001 | *** |
| Δ Age: Δ LYg | -0.36 | 0.05 | -7.56 | <0.001 | *** |
| Δ Final health state: Δ LYg | 3.23 | 0.55 | 5.87 | <0.001 | *** |
| Δ Patients treated: Δ LYg | -0.38 | 0.04 | -8.45 | <0.001 | *** |
| Δ Age:Agent | 1.19 | 0.72 | 1.64 | 0.101 | |
| Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+' | | | | | |

Appendix 9.5: Latent class multinomial logistic coefficients and differences by questionnaire format

| Attribute | DCE | | | CSPC | | | Difference | | | | |
|------------|-------------|-----------|----------|-------------|-----------|----------|------------|--------|--------|------------|-----|
| | Coefficient | Std Error | P[Z >z] | Coefficient | Std Error | P[Z >z] | (DCE-CSPC) | L95CI | U95CI | Adjusted-p | Sig |
| LYg | 0.283 | 0.061 | 0.000 | 0.116 | 0.017 | 0.000 | 0.167 | 0.081 | 0.253 | 0.004 | ** |
| Age 10 | 1.981 | 0.861 | 0.021 | 0.530 | 0.082 | 0.000 | 1.451 | -0.077 | 2.979 | 0.211 | |
| Age 70 | -1.571 | 0.589 | 0.008 | -0.704 | 0.112 | 0.000 | -0.867 | -1.805 | 0.070 | 0.230 | |
| U0 0.1 | -2.329 | 0.759 | 0.002 | -0.353 | 0.089 | 0.000 | -1.976 | -3.291 | -0.660 | 0.031 | * |
| U0 0.9 | 2.094 | 0.472 | 0.000 | 0.498 | 0.137 | 0.000 | 1.596 | 0.939 | 2.252 | <0.001 | *** |
| LE 1m | -1.317 | 0.355 | 0.000 | -0.281 | 0.069 | 0.000 | -1.036 | -1.599 | -0.474 | 0.007 | ** |
| LE 10yrs | 0.531 | 0.223 | 0.017 | -0.027 | 0.039 | 0.479 | 0.558 | 0.195 | 0.921 | 0.025 | * |
| U1 0.1 | 0.845 | 0.258 | 0.001 | 0.229 | 0.130 | 0.078 | 0.616 | 0.365 | 0.868 | <0.001 | *** |
| U1 0.9 | -1.863 | 0.642 | 0.004 | -0.209 | 0.101 | 0.038 | -1.654 | -2.716 | -0.592 | 0.023 | * |
| nPats 100 | -0.031 | 0.148 | 0.834 | 0.069 | 0.054 | 0.204 | -0.100 | -0.285 | 0.085 | 0.295 | |
| nPats 5000 | 1.397 | 0.490 | 0.004 | 0.134 | 0.056 | 0.016 | 1.262 | 0.409 | 2.115 | 0.035 | * |
| (1-U0):U1 | 0.050 | 0.029 | 0.083 | 3.063 | 0.456 | 0.000 | -3.013 | -2.174 | -3.852 | <0.001 | *** |
| LYg:Age | -0.371 | 0.139 | 0.008 | | | | -0.371 | -0.644 | -0.097 | 0.024 | * |
| LYg:U0 | -0.013 | 0.006 | 0.034 | | | | -0.013 | -0.024 | -0.001 | 0.153 | |
| LYg:LE | -0.045 | 0.057 | 0.438 | | | | -0.045 | -0.157 | 0.068 | 0.295 | |
| LYg:U1 | -0.017 | 0.012 | 0.154 | -0.049 | 0.017 | 0.000 | 0.032 | 0.043 | 0.021 | <0.001 | *** |
| LYg:nPats | 8.752 | 2.080 | 0.000 | | | | 8.752 | 4.667 | 12.836 | 0.001 | ** |

Significance codes: <0.001= '***' <0.01= '**' <0.05= '*' <0.10= '+'

The DCE model is the same model estimated in the primary analysis. The initial CSPC model included all dummy-coded main effects and continuous interactions. Parameters that were not significant at a 0.10 threshold in any of the latent classes were excluded and the model was re-estimated. LYg=individual life years gained; U0=initial health state; U1=final health state; LE=initial life expectancy; nPats=total patients treated; ASC=alternative-specific constant (intercept); L95CI/U95CI=lower/upper 95% confidence interval.

Chapter 10:

Discussion and concluding remarks

The thesis started from the position that greater explicitness in the criteria by which competing claims to limited healthcare resources are prioritised is, on the whole, desirable. Explicitness, in the sense that prioritising decisions are made on the basis of clearly defined objectives and criteria, is argued to promote a more informed citizenry, more accountable decision makers, greater opportunities for evaluation and improvement, greater trust in the priority setting process, and – ultimately – better decisions.

The theoretical QALY maximisation framework has the advantage of an explicit definition of value as the sum of individual health-related utilities, but it has faced criticism over the narrowness of its definition of well-being and its presumption of distributive neutrality. As a result, most jurisdictions that use some form of QALY maximisation, including the UK, Canada and Australia, implicitly consider equity alongside efficiency in priority setting decisions (see for example National Institute for Health and Clinical Excellence 2008; Canadian Agency For Drugs and Technologies in Health 2013; pan-Canadian Oncology Drug Review 2011; Australian Government Department of Health and Ageing 2008). Such operational frameworks, though, have tended to avoid explicitly defining a set of relevant equity concerns and their relative weights.

Advocates for greater explicitness in priority setting argue that the solution is not to accept efficiency and equity as inherently separate and incompatible issues, but rather to incorporate more of the elements that contribute to societal value and well-being within an explicit framework. Indeed, the original motivation for the QALY within the extra-welfarist approach was to move beyond individual utility in priority setting decisions, and

Culyer (1989) suggested that combining health outcomes and distributional concerns into a single outcome measure – such as the equity-weighted QALY – could achieve an explicit integration of equity and efficiency in healthcare priority setting. A key issue in developing such a comprehensive outcome measure, though, is the identification and relative weighting of relevant equity concerns.

At one extreme, Brouwer et al. (2008) suggested that decision makers should define well-being in terms of the preferences they believe society *ought* to hold. This approach has the advantage of professional judgement and perceived objectivity. As a number of authors have noted, however, there is no objective basis for preferring one (Pareto optimal) allocation of healthcare resources over another; it is an inherently subjective value judgement. In their view, the importance of different perspectives in arriving at an optimal and legitimate allocation should be acknowledged, not avoided, leading to democratic or Communitarian approaches which allow for the community to define its own equity weights, and measure societal value on the basis of how well a program satisfies community preferences.

Within this context, the thesis made a number of specific contributions that may inform healthcare priority setting:

- It reviewed patient and program characteristics that were potentially relevant to the allocation of healthcare resources, and through an application of empirical ethics narrowed these potential characteristics to those that had evidence of public support and a defensible ethical justification; that is, were *relevant* and *fair*.
- It reviewed different stated preference methods and identified two methods that appeared particularly suited for eliciting preferences in a healthcare context: discrete choice experiments (DCEs) and constant-sum paired comparisons (CSPCs), and compared their response characteristics to identify a preferred method for the elicitation of societal preferences in healthcare.
- It elicited preferences over the allocation of societal healthcare resources in order to estimate the equity-efficiency trade-off associated with the

attributes identified in the empirical ethics review, and compared preferences between the public and a sample of decision-making agents.

There were several unique elements among these contributions that warrant highlighting. First, this study was the first to the author's knowledge to use an empirical ethics approach in identifying the attributes to be included in a stated preference elicitation. Most elicitations in this area have used focus groups or literature review to identify potentially relevant attributes, but have not included any process to ensure that the preferences being elicited were, in some sense, fair. The process used here began with a literature review to identify attributes that had empirical evidence of public support, but also required those attributes to be consistent with some coherent theory of distributive justice. In addition to ensuring that preferences are elicited over attributes that are fair as well as relevant, an empirical ethics approach may contribute to standardising the set of attributes over which preferences for the allocation of healthcare resources are elicited. Given the importance of context in stated preference elicitations, standardising the identification of attributes would help make the results of elicitations in this area more comparable. However, it is important to acknowledge the subjectivity inherent in this approach, both in identifying appropriate principles of justice and in interpreting the empirical evidence with respect to these principles. This subjectivity may limit reproducibility, but it can be seen as an essential characteristic of the empirical ethics approach that was applied here, as more systematic approaches may lead to a greater emphasis on empirical observation at the expense of ethical judgement.

Second was the use of CSPC to elicit societal preferences, and to the author's knowledge, the first head-to-head comparison of CSPC and DCE for eliciting societal preferences. A CSPC allocation task makes it clear that giving more resources to one group means that the other must necessarily receive less. This trade-off can be obscured in a discrete choice task, where it may not be clear that choosing one group implies that the other will receive no resources. The incidence of non-compensatory decision making among DCE and CSPC respondents, tested as part of the head-to-head comparison, appeared to support the characterisation of the CSPC as a more reflective task. The comparison also

suggested that relative to the DCE, the CSPC was associated with greater concern for the less well-off group and less consistency with the principles of QALY maximisation. The implications of this finding are discussed in more detail in section 10.3 below. Despite its theoretical advantages, Table 4.2 suggested that CSPC has not been used as often as DCE for eliciting societal preferences. Among the studies that have used CSPC, a slight majority have taken a categorical approach to analysing the responses, simplifying the continuous allocations into discrete categories. This neglects the cardinal preference data that is a statistical advantage of the CSPC method, and makes it difficult to consider the impact of multiple attributes simultaneously. The CSPC administered here also appeared to be unique in dynamically linking some attribute levels to the relative budget allocation, further highlighting the trade-offs associated with prioritising one group over the other. This dynamic linkage allowed the elicitation to test for preferences for equality in access (number of patients treated) or outcomes (aggregate QALYs gained) in addition to the more straightforward equality in resources. No other CSPC elicitation has included attributes that varied with the relative budget share, although Linley and Hughes (2012) took a direct approach and asked respondents how many patients from each of two equally-sized groups they would prefer to treat, skipping the intermediate step of allocating a budget. Treating more patients from one group meant that fewer patients from the other group could be treated, emphasising the trade-off, but the relative allocation of patients was not linked to an aggregate outcome as it was here. Finally, it is worth noting that only one other CSPC elicitation (Linley & Hughes 2012) took a representative sample, as was used here.

Third was the use of a latent class approach to model the DCE responses. As shown in Table 4.2, statistical models of DCE and CSPC panel data have most often used a random effects specification. With this specification, however, individual preferences differ only because each individual is an independent draw from a pre-specified random distribution (Morey & Greer Rossmann 2003). In contrast, a latent class approach derives preferences from individual choice behaviours and observed or unobserved respondent characteristics. It also allows respondents to be assigned to latent classes on the basis of their choice

behaviours rather than a deterministic characteristic. For example, the latent class analysis of DCE responses suggested that agents were substantially more likely to belong to one particular class, but it allowed for some agents to have a different set of preferences and not to be defined by their status as an agent. The need to pre-define an arbitrary number of latent classes, and the assumption that preferences are homogeneous within each class, are limitations of the approach, but relative to a continuous distribution of preferences from a random parameters model, this simplification improves the interpretability and salience of the estimates. In particular, it allows the interpretation of the different latent classes in terms of the characteristics of their members, in a way that would not be possible with a random effects specification. It was suggested in section 8.5, for example, that the defining latent characteristic among the DCE respondents might have been related to the axiomatic quality of their preference formation. This may have implications for how societal preferences are elicited and interpreted, and is discussed in more detail below.

The implications of these contributions are discussed below. Section 10.1 discusses the evidence from the DCE and CSPC elicitations for an equity-efficiency trade-off in societal preferences, and section 10.2 discusses this evidence in the context of other recent societal elicitation in healthcare. Section 10.3 summarises the response behaviours of the DCE and CSPC methods with respect to identifying a preferred method for future elicitation in this area. Section 10.4 discusses the implications of the overall results for healthcare priority setting, while section 10.5 outlines some of the caveats and challenges to incorporating societal preferences into this process. Finally, section 10.6 offers some concluding remarks.

10.1 An equity-efficiency trade-off?

The results showed little support for the principles of strict QALY maximisation as a societal decision rule, as fewer than 5 percent of all respondents always prioritised the alternative that maximised aggregate QALYs gained, and decision making agents were no more likely than the general public to make such choices. It was difficult, however, to define a threshold at which

one would be confident in accepting or rejecting the relevance of this decision rule. The imperfect and probabilistic nature of preference elicitation under the assumptions of random utility theory means that even if there was universal support for strict QALY maximisation, it is unlikely that every respondent would be observed prioritising the QALY maximising alternative with every choice. As such, a threshold of perfect and unerring consistency is unrealistic. Bryan et al. (2002) suggested that a majority of respondents maximising QALYs gained over a majority of their choices may be sufficient to indicate support for QALY maximisation. By this standard support for QALY maximisation was stronger, as 75 percent of all DCE respondents chose the QALY maximising alternative in at least half of their choices, although only 41 percent of CSPC respondents allocated the majority of the budget to the QALY maximising alternative in at least half of their choices. As in any stated preference survey, though, it is not possible to say with certainty *why* a particular alternative was chosen, so it is important to recognise that QALY maximising alternatives may have been chosen for reasons unrelated to aggregate QALY gains.

Despite the lack of strict QALY maximising behaviour, the rankings of the DCE and CSPC scenarios showed that larger aggregate QALY gains tended to be ranked more favourably than those with smaller gains. However, the rankings also suggested an equity-efficiency trade-off, as respondents appeared willing to prioritise relatively small aggregate QALY gains to preferred patient groups over larger QALY gains to less preferred groups. In particular, younger patient groups were consistently preferred to older patient groups, even when the older patients had the potential to gain a greater number of QALYs.

The idea of an equity-efficiency trade-off was further supported by the marginal analyses, which found a statistically significant willingness to sacrifice life year gains in order to prioritise younger patients, and those who would finish treatment in better final health states. The CSPC also found a preference for patients with the worst initial quality-of-life, while the DCE found a preference for patients with longer initial life expectancies – even after controlling for potential health gain – and for larger patient groups. The willingness to forego individual life year gains in order to prioritise larger patient groups suggested a

preference for smaller individual gains distributed over more beneficiaries to larger individual gains concentrated amongst a smaller number of beneficiaries.

Attribute main effects were significant in both the DCE and CSPC models, even after interactions with life year gains were included in the value functions. Some recent elicitations of societal preferences over the allocation of healthcare, including Norman et al. (2013) and Lancsar et al. (2011), did not include independent main effects in their value functions, consistent with a consequentialist view that healthcare is not valued for its own sake but rather for the health outcomes it delivers (Mooney 1998a). However, the significant main effects observed here suggested the possibility of welfare gains associated with treating particular patient groups, even in the absence of health gains. For example, the positive and significant welfare effects associated with treating younger patients, independent of their potential life year gains, suggested that society may derive value from seeing young patients receive care, even if that care does not lead to improved outcomes. Such welfare effects would be consistent with arguments that society may desire a healthcare system that provides for aspects such as compassion, respect for dignity, and maintenance of hope, in addition to health gains (Donaldson & Shackley 1997; Wiseman 1997; Mooney 1998a; Salkeld 1998). This possibility was not specifically tested here, but future research would be useful in identifying specific non-health factors that might be associated with societal welfare gains.

10.2 Comparison with other societal preference elicitations

There were some similarities between the methods and results reported here and other recent societal elicitations of preferences over societal healthcare resources, but the overall body of societal preference research in healthcare is characterised by its heterogeneity rather than its consistency. As noted, other elicitations of societal preferences in healthcare have identified different sets of relevant attributes, using different methods and different inclusion or exclusion criteria. Green and Gerard (2009), for example, acknowledged that empirical evidence appears to suggest that the public supports prioritising younger patients, but excluded this factor from their societal DCE primarily on the grounds that

NICE does not support age as an independent factor in priority setting.¹⁷ Conversely, Norman et al. (2013) included social role, for which the empirical ethics review found a utilitarian justification but little evidence of public support, personal responsibility, for which there was evidence of strong public support but limited ethical justification, and patient gender, for which there appeared to be neither public support nor a clear ethical justification.

There is nothing within democratic or Communitarian approaches that require community preferences to meet any ethical standard. But if one of the objectives of societal participation in priority-setting is to improve the moral legitimacy of the allocation of resources, it is difficult to accept that this objective would be furthered by incorporating ethically questionable preferences. Given the importance of context in stated preference elicitation, the inclusion or exclusion of different attributes may also lead to different trade-offs. It is important to recognise, therefore, that the process of identifying a set of relevant attributes may be as important as the elicitation methods themselves in arriving at a meaningful set of preference estimates, and a more standardised approach may be necessary before elicited preferences can be used for policy in the form of equity weights. Relative to using literature review or focus groups as a basis for identifying attributes relevant to the allocation of societal resources, an empirical ethics approach may help standardise the attributes included in future elicitation and ensure that these attributes are fair as well as relevant. However, the subjectivity in interpreting the ethical justifications for different attributes, as well as in identifying 'defensible' theories of justice, must be acknowledged and may limit the degree to which an empirical ethics approach can in itself contribute to this more standardised approach. There is no universal theory of justice, so it is necessary to judge the applicability and appropriateness of different theories of justice in the context of allocating healthcare. Likewise, the degree to which different patient or program characteristics are consistent with these theories is

¹⁷ NICE guidelines state that "patients should not be denied, or have restricted access to, NHS treatment simply because of their age." (National Institute for Health and Clinical Excellence 2008) Age is relevant, though, when it is a predictor of treatment outcomes or closely associated with some aspect of a patient's health status or likelihood of adverse events. The UK Equality Act also prohibits discrimination on the basis of age in the provision of public services, including healthcare (Carruthers & Ormondroyd 2009).

also a matter of subjective interpretation. However, as Richardson (2002) stresses, “an integral part of empirical ethics should be an acceptance of the fact that argument and evidence are fallible and the conclusions are tenuous and more or less strongly supported in some contexts than others.”

A summary of recent preference elicitations in healthcare is shown in Appendix 10.1, along with a summary of the results from the DCE and CSPP administered as part of the thesis. The attributes are ordered by the frequency with which they were included in the different studies. Life years or QALYs gained was the most commonly included attribute, followed by patient age, life expectancy, and initial and final quality of life. The remaining attributes were included in only two or three studies each. Greater life year or QALY gains were preferred in all studies that included them, and Bryan et al. (2002), Green and Gerard (2009), Koopmanschap et al. (2010), and Lancsar et al. (2011) concluded that their results were consistent with a QALY maximising decision rule. Similar studies by Schwappach (2003), Dolan et al. (2008), Linley and Hughes (2012), and Norman et al. (2013), though, found a willingness to forego potential health gains to prioritise specific patient characteristics and concluded that respondent preferences were not consistent with QALY maximisation.

Among the other elicitations, younger patients were consistently preferred to older patients. It is worth highlighting that this preference was observed across a number of different countries, including the UK (Ratcliffe 2000; Dolan & Tsuchiya 2005; Dolan et al. 2008; Baker et al. 2010; Petrou et al. 2013), Germany (Schwappach 2003), Hong Kong (Chan et al. 2006), and in this study, Canada. However, two other UK studies found no significant preferences over age (Lancsar et al. 2011; Linley & Hughes 2012), and a German study found a non-linear preference that peaked at middle age and declined over older and younger patients (Diederich et al. 2012). There was also broad agreement across the studies in favour of prioritising patients who would finish treatment in better final health states and those with lower levels of personal responsibility for their illness, although Schwappach (2003) found unexpected support for prioritising patients with greater responsibility for their disease. Support for prioritising patients in the worst initial health states was mixed. Most studies, including the current CSPP, found support for prioritising patients in poorer initial health

states, but Shah et al. (2012) found the reverse, and Dolan and Tsuchiya (2005), Lancsar et al. (2011), and the current DCE found no significant preference. Overall, with the exception of a consistent preference for greater life year or QALY gains, there was some degree of heterogeneity, either in a conflicting direction of preference, or no statistically significant preference, in each of the other commonly included attributes.

As noted earlier, the primary DCE and CSPC elicitations in the thesis were conducted in a cancer context. This did not appear to have affected the interpretability of the results relative to the other studies included in Appendix 10.1, a handful of which were also conducted in specific disease contexts, including liver transplant, orphan diseases, and cancer.

10.3 Choosing between the DCE and CSPC

The DCE and CSPC elicitations revealed the tremendous potential of stated preference methods, as respondents were able to make remarkably coherent choices over very complex sets of attributes and trade-offs – even in the less intuitive CSPC – with very minimal instruction. Based on questionnaire completion rates, completion times, and difficulty ratings, the DCE appeared to be the more straightforward task, although the more competitive nature of the DCE also appeared to be associated with a greater incidence of non-compensatory decision making, as respondents were significantly more likely to hold a dominant preference for a single attribute. The superior completion rates and high preference confidence ratings in the DCE appeared to reject Swallow et al.'s (2001) contention that respondents may be reluctant to complete dichotomous preference elicitations over highly emotive issues. The results of the elicitations also appeared to discount the hypothesis of a prominence effect around the number of patients treated attribute in the CSPC, as DCE respondents gave more weight to this attribute in their decisions, consistent with the more competitive and quantitative tendencies of the task. The emphasis in the CSPC on individual over aggregate QALY gains was also consistent with suggestions that respondents to complex societal stated preference elicitations may tend to reduce the abstract, macro-level allocation problems to more

comprehensible two-person analogies. This may have been exacerbated by the individual-level presentation of the QALY graphs, although it is notable that a corresponding emphasis was not observed in the DCE, where respondents saw the same graphs but put relatively more weight on aggregate QALY gains. Finally, one cannot discount the possibility that the range of the patients treated attribute was simply too narrow to distinguish preferences for different levels (Kjær 2005). In the absence of the theorised advantages of CSPC, the greater completion rate and slightly more favourable difficulty rating of the DCE appeared to give it an advantage in eliciting societal preferences.

However, a notable difference between the DCE and CSPC was in the greater willingness of CSPC respondents to prioritise the group with the poorer health prospects in the test of non-satiation. CSPC respondents also consistently prioritised alternatives associated with fewer individual and aggregate QALY gains. It was hypothesised that this may reflect a compassion bias in the CSPC, inherent to the nature of the task, which required respondents to consider how much of the budget, if any, to reserve for the less preferred group in each task. CSPC respondents could see that as they allocated more of the budget to one group, fewer patients in the other group could be treated. DCE respondents faced this same trade-off, as choosing one group meant that none of the patients in the less preferred alternative would be treated, but the trade-off was not made as clear as in the CSPC, and they may not have had to confront fully the consequences of their choices.

The possibility of such a tendency in the CSPC, which appeared to encourage a relatively higher proportion of prioritisation choices that might be deemed irrational by economic theory, may be linked to the issue of hypothetical bias in stated preference elicitations. DCE respondents were more consistent with economic theory in more often choosing QALY maximising alternatives and the dominant alternative in the test of non-satiation, but this consistency may be in part an artefact of the competitive focus of the task and might not be observed in real-life choices. The arguably less rational CSPC choices might be more reflective of how respondents would choose in a real-life situation. Indeed, the results of the CSPC, which suggested preferences for patients in poorer initial health states, and those that could be returned to better final health states,

appeared more consistent with expectations from the empirical ethics review than the results from the DCE, which found no significant preference over those attributes. Note, though, that the results of the empirical ethics review did not generally account for opportunity costs or the relative strength of preferences, and so should not be interpreted as a gold-standard or assigned any normative qualities. Overall, the differences between the DCE and CSPC were suggestive of a violation of the conventional assumption of procedural variance between stated preference methods – the observed preferences appeared to be systematically influenced by how they were elicited. To the extent that responses to the DCE may have reflected a greater degree of hypothetical bias, the CSPC may be a more appropriate stated preference method for eliciting societal preferences over emotive issues such as the allocation of societal healthcare resources. More research will be required to verify the existence of a relative tendency towards compassion in the CSPC, and to establish which format produces a result that is more consistent with a reflective equilibrium.

10.4 Implications for healthcare policy

The results of the DCE and CSPC elicitations conducted as part of the thesis, as well as the other recent elicitations, suggested that there were statistically significant welfare effects associated with attributes that are not generally considered within the theoretical QALY maximising framework. In light of these effects, the aggregate value that society derives from healthcare might be improved by giving explicit weight to attributes such as the age of the patient and their expected final health state in priority setting decisions in an equity-weighted QALY. Note, however, that even if it had been found that societal preferences were entirely consistent with strict QALY maximisation, it can be argued that there was additional value in an inclusive approach. As argued in Chapter 2, societal participation can enhance the moral legitimacy of the resulting decision rule, and similarly improve public trust in the priority setting process.

In this context, though, it is important to recognise that that equity weighting simply *redistributes* healthcare resources. The sum of equity weighted

QALYs must equal the sum of unweighted QALYs, and for each patient that receives higher priority with equity weighting another must necessarily receive lower priority (Ham & Coulter 2001; Wailoo et al. 2009). Implicit equity weights may obscure this reality and make it easier for decision makers to implement priority setting decisions, but they raise the spectre of Fleck's (1992) invisible class of 'others,' as less preferred patients may not realise that they have been given lower priority relative to others. Under implicit weighting, different decision makers may also assign different weights to different patient characteristics, leading to inconsistent decisions that may jeopardise public trust in the priority setting process.

Interestingly, the DCE analysis found a significant welfare effect associated with initial life expectancy, but contrary to a NICE supplementary advice that advised giving greater weight to health benefits to patients with less than 24 months life expectancy (National Institute for Health and Clinical Excellence 2009), the DCE result suggested that prioritising patients with the shortest life expectancy was associated with a welfare *loss*, even after controlling for potential health gain. A similar preference for patients with greater untreated life expectancy was found by Schwappach (2003), Lancsar et al. (2011) and Shah et al. (2012). This inconsistency may be explained by NICE's acknowledgement that there was no consideration of the opportunity cost of giving greater weight to patients at the end of life. Indeed, the advice appeared to be based largely on the fact that 63 percent of stakeholder respondents supported the proposition. This highlights the importance of giving explicit consideration to the relative strength of preferences in priority setting decisions, as it appears that accounting for the trade-offs with other factors might have led to a different advice regarding priority for patients at the end of life.

Together, these results suggest that giving greater priority to the healthcare claims of younger patients and those that would finish treatment in better health states may enhance overall societal welfare. Additional weight to such patients in an equity-weighted QALY would prioritise health gains accruing to particular patients in priority setting decisions, but given the suggestion above of societal value associated with non-health outcomes, priority may need to extend to the patients themselves, and not just their health gains.

Future research is required to understand to what extent society would be willing to trade-off health gains for non-health outcomes, and how to incorporate these preferences into priority setting criteria. This issue is similar to the challenge of measuring value in palliative care, where health economists have struggled to measure the value of care that is more often associated with non-health outcomes such as dignity and respect than conventional QALY gains (Normand 2009).

Before implementing a policy of greater priority for specific patient groups, it is important to recognise that there are legislative prohibitions on discriminating between citizens on the basis of personal characteristics. Despite clear preferences for greater priority for younger patients and for those that would finish treatment in better final health states, statutes such as the Canadian Human Rights Act (Anon 1985) and the UK Equality Act (Anon 2010) prohibit discrimination on the basis of age and disability. Similar to the role of the theories of justice discussed in Chapter 3, such legislation ensures that the basis of societal resource allocations are just and do not reflect irrational or perverse preferences. There may still be some scope within such legislation, though, for prioritising on the basis of such attributes. For example, a prohibition on discriminating on the basis of disability does not preclude prioritising the more severely ill in a triage setting. The balance between protecting the rights of the minority while reflecting the preferences of society is a complex issue, but societal preference elicitation such as these can help inform such deliberations.

There are also a number of methodological challenges to using the equity-weighted QALY in priority setting, including how to accommodate changing patient characteristics over time and technical challenges to incorporating these weights in economic models (Wailoo et al. 2009; Baker et al. 2010). The next section discusses a number of other methodological issues that must be resolved before societal preferences can be used to inform healthcare priority setting.

10.5 Methodological challenges to incorporating societal preferences into healthcare priority setting and suggestions for future research

There are a number of substantial methodological challenges to incorporating societal preferences into an explicit priority setting framework such as the equity-weighted QALY. This includes the appropriate method for calculating equity weights themselves. For example, Dolan et al. (2008) and Lancsar et al. (2011) estimated societal equity weights for specific patient characteristics, but they used different methods to calculate these weights and they reached different conclusions. Dolan et al. calculated that health gains to children had a statistically significant equity weight of 1.8 relative to adults, while Lancsar et al. calculated that equity weights for younger patients relative to a 40 year-old ranged from 0.98 to 1.02 and were not statistically significant. These differences may reflect differences in the respective methodologies of the two studies, but the more general issue of how to reconcile discordant societal preferences is discussed below, along with the question of how to elicit representative preferences, and the limited public desire to participate in the priority setting process.

10.5.1 Aggregating heterogeneous societal preferences

Within the DCE, the two latent classes identified held statistically significant but offsetting preferences over initial and final health states and the number of patients treated. These results highlighted the value of latent class modelling, but also highlighted a fundamental challenge to incorporating societal preferences into societal decision-making: respondents in the two latent classes held statistically significant but opposing preferences for specific patient characteristics that effectively cancelled each other out, resulting in an insignificant overall preference. It is not clear how such opposing preferences can or should be reconciled. The combined, statistically insignificant results did not represent the significant preferences of either class, but basing decisions on the preferences of just one of the classes effectively imposes their preferences on the other class. In this case, the problem was compounded by the fact that the probability of being in either of the two classes was roughly equal, meaning there was no scope for an appeal to the ‘will of the majority.’

This dilemma recalls Arrow's (1963) impossibility theorem, which showed that there is no process by which individual preferences can be aggregated in a way that satisfies a relatively weak set of axioms. These axioms included an unrestricted domain of alternatives, unanimity or weak Pareto (if all individuals prefer x to y , then society must prefer x to y), non-dictatorship (no individual can impose their preferences on society), transitivity (if $x \succ y$ and $y \succ z$, then $x \succ z$), and independence from irrelevant alternatives (the relative ordering of x and y should not depend on the inclusion or exclusion of z) (Mullen & Spurgeon 1999; Mueller 2003). Society can only escape this quandary by relaxing one of these axioms.

Perhaps the most obvious solution may be to relax the non-dictatorship axiom and allow some expert or impartial party to reconcile the opposing preferences. As Mueller (2003) noted, there is nothing unusual or irrational about allowing small groups to make decisions on behalf of a community or organization. Indeed, this logic is the basis of the decision maker perspective, and by extension, QALY maximisation. However, it effectively allows decision makers to decide which societal preferences count, again defeating the spirit and objective of a democratic or Communitarian approach.

Mullen and Spurgeon (1999) suggested that a less dictatorial approach may be to give different members of society different weights in the aggregation of preferences, presumably estimated through a process similar to the estimation of allocative preferences described herein. However, this would seem merely to shift the problem from reconciling opposing preferences over the weight to give to different patient characteristics to reconciling opposing preferences over the weight to give different citizens. In addition, there is nothing in a 'citizen-weighted' solution that would guarantee that the resulting weighted preferences for different patient characteristics would not still be similarly offsetting.

In light of these shortcomings, Mueller (2003) suggested that a more pragmatic solution was not to relax the non-dictatorship axiom, but rather the transitivity axiom. Transitivity is fundamental to many aspects of economic theory and is critical in avoiding the problems of cyclical preferences (i.e. if $x \succ y$ and $y \succ z$, but $z \succ x$, individuals can get caught in a cycle of voluntary trades that leave them worse off than their original state (Feldman & Serrano 2006)).

However, Mueller noted that the enduring popularity of arbitrary processes such as coin flips or drawing straws to resolve conflict suggested that the perceived impartiality and fairness of such a solution may be more a fundamental requirement of societal decision making than establishing a transitive ranking of all alternatives. A good example is a coin flip to settle an election where two candidates received an equal number of votes: a transitive ranking of candidates may not have been established, but all sides can accept the result as fair.

In the context of the offsetting preferences observed here, relaxing the transitivity axiom would mean accepting equal weights over the conflicted levels of an attribute as an arbitrary resolution of the conflict. Significant preferences for particular levels of an attribute would have no bearing on prioritising decisions, but as each group's preferred level would have an equal opportunity of being prioritised this may be acceptable as a fair solution to an otherwise difficult quandary. Note that this is not the same as *a priori* omitting an attribute from consideration – such laundering perverse preferences in the empirical ethics review – as relaxing transitivity still allows each individual to express a preference, even if aggregation may ultimately render that preference insignificant in the distribution of resources.

The dilemma of how to aggregate preferences persists, though, if some citizens will not accept arbitrarily equal weights as a fair solution. Some individuals or groups may adamantly resist equal weights over different levels of an attribute that they feel embodies a fundamental or protected value. For example, supporters of an absolute age cut-off on health expenditures may strongly resist equal priority over age. It is irreconcilable dilemmas of this sort that advocates of a more implicit approach point to as justification for a more deliberative, political process (Klein 1997; Hunter 2001). This leads back to Arrow's quandary and suggests, somewhat perversely, that the greatest challenging to incorporating the strength of individual preferences into societal decision-making may be the very strength of some preferences. Indeed, the nature of individual preferences likely means that complete and transitive rankings of preferences may only be achievable in very limited circumstances (Sen 1992; Mooney 1998b).

10.5.2 Eliciting reliable preferences

Similar to the question of which stated preference format is ‘best’ for eliciting preferences is the question of whether preferences should be elicited from citizens ‘off the top of their heads,’ or following some deliberative process (Dolan et al. 2008). The results presented in this thesis were based on ‘top of the head’ elicitations, with little opportunity for respondents to reflect on their preferences. There was, though, anecdotal evidence of respondents changing their earlier answers as they progressed through the questionnaires and their understanding of the issues and trade-offs evolved.¹⁸ Although preferences elicited in this manner are generally held to be representative, it is not clear that representativeness in this sense should trump the possibility that more considered and reliable preferences may emerge from a deliberative exercise. Gregory, Lichtenstein and Slovic (1993) argued that stated preference elicitation should be seen as a process of preference construction rather than a neutral process of preference discovery, and suggested that a more deliberate process may improve the result. As Hausman and McPherson (2006) argued, well-being can only be equated with the satisfaction of preferences if those preferences are well-informed and well-considered. There is evidence that a deliberative process can change an individual’s stated preferences (Abelson, Eyles, et al. 2003), and Dolan et al. (2008) acknowledged that if these changes stem from better knowledge about one’s own preferences and those of others in the community, then a more deliberative process is probably superior. However, they also noted that if these changes stem from a social desirability bias or ‘bandwagon effect,’ then ‘top of the head’ preferences may be preferred. In addition, Abelson et al. (2003) note that although deliberation has come to be understood as requiring some interaction amongst a group, as a process of weighing evidence and reasons there is no reason that it cannot be seen as an individual activity. That is, well-considered preferences do not necessarily have to derive from a group activity. A better understanding of how, and more importantly, why, preferences change following a deliberative exercise, and how this might be different in an individual

¹⁸ A number of respondents interviewed after completing the pilot survey reported that they initially prioritised the younger age group in each choice, but changed these answers as they began to better recognise the trade-offs with the other attributes.

and a group context, will be essential to ensuring that the allocation of healthcare resources is based on well-informed preferences, and this represents a critical area of future research.

Similarly, there are concerns over the ability of respondents to complex stated preference elicitation, particularly in the context of healthcare, to understand and process the information they are given (Ryan et al. 2001; Dolan et al. 2008; Baker et al. 2010). These concerns are shared here, as it was difficult to be confident that the respondents to the DCE and CSPP fully understood the nuances of the attributes they were asked to consider, such as a patient's experience at different levels of utility, or the concept of the QALY. These are complex concepts to explain, particularly with only a brief online description. A face-to-face elicitation format may have been more effective at helping respondents develop a full understanding of these concepts. However, as decision-making agents were presumably more familiar with many of these concepts, the lack of significant differences between the preferences of agents and the general public was somewhat reassuring in this regard. It suggested that the public had at least a comparable understanding of these attributes relative to that of the likely somewhat more sophisticated agents.

10.5.3 Public involvement in priority setting

The non-significant differences between the preferences of agents and the general public also call into question the necessity of societal participation in healthcare priority setting: if agents generally hold the same preferences as society at large, why devote the time and expense required to involve the public? Part of the answer lies in the fact that one can only establish the representativeness of agent preferences by also asking the public about their preferences. However, the thesis has also offered a number of theoretical advantages associated with more inclusive and participatory approaches to priority setting, including the promotion of legitimacy and trust in the process, and the ethical importance of 'universality of inclusion.' It is also possible that there may be 'procedural utility' associated with more inclusive or participatory approaches to priority setting. Procedural utility suggests that society may derive value from the process by which a decision is made, independent of the

outcomes associated with that decision (Frey & Stutzer 2005; Dolan et al. 2007). An instrumentalist interpretation of procedural utility suggests that people may value a particular process because they believe it will arrive at an outcome with which they will be more satisfied, which is little different than a consequentialist view. Non-consequentialist interpretations of procedural utility, though, hold that people may value a particular process for its own sake, independent of its outcomes, because such a process may be more consistent with valued principles. This interpretation was supported by Dolan et al. (2007), who found that among other factors, the public appeared to value what they referred to as ‘voice,’ or the opportunity for affected parties to participate in decision processes.

However, the attitudinal results presented in Chapter 6 suggested that only about half of the general public respondents would prefer to see the public have a role in priority setting decisions, and a similar proportion indicated that they would be uncomfortable having their preferences used to set priorities. This result was consistent with Lomas’ (1997) characterisation of citizens as “reluctant rationers,” as well as with other similar findings. This general reluctance to participate in societal priority setting leads to questions about the representativeness of those members of the public who *are* willing to participate, as to some extent their very willingness to participate makes them unrepresentative of the larger community (Mullen & Spurgeon 1999). Those who are willing to participate may be more likely, or at least perceived to be more likely, to have a specific motive or agenda, weakening the moral legitimacy derived from public participation.

As discussed in Chapter 2, Fleck’s (1992) resolution to this issue was to emphasise the obligations, in addition to the rights, of citizens in a democracy, and he more or less endorsed coercing citizens into participation, presumably similar to compulsory voting laws in many jurisdictions. However, involuntary participation clearly has its own drawbacks, most particularly around the effort such participants might be likely to devote to the task. Studies of compulsory voting have found higher rates of invalid ballots and voters simply choosing the name at the top of the ballot (Jackman 2001). As the time and cognitive effort required to participate in a stated preference elicitation would seem to exceed that of voting, it is likely that there would be a substantial proportion of invalid

responses associated with compulsory participation in a priority setting process, which may undermine the result and the overall objective. Coercion would also seem to negate any procedural utility gains that might be associated with more inclusive processes, and many participants may in fact experience negative procedural utility as a result of this coercion. This reluctance appears to represent the greatest barrier to incorporating societal preferences into healthcare priority setting. Litva et al. (2002) noted that a process where all citizens have an genuine opportunity to participate in setting system-level priorities may generate greater support and participation than indirect consultations, but the benefits of more inclusive priority setting approaches may remain theoretical unless a broad segment of society chooses to participate.

10.6 Concluding remarks

The stated preference elicitation administered here appeared to reject strict QALY maximisation as a societal decision rule for allocating healthcare resources, and instead suggested a clear equity-efficiency trade-off in societal preferences. Strict QALY maximisation is, admittedly, something of a straw man, as few, if any, jurisdictions actually adhere to this rule in societal priority setting decisions, and most include some implicit consideration of equity factors. However, implicit consideration of equity was argued to be insufficient, as it fails to account for the relative strength of the equity-efficiency trade-off for different characteristics, and has the potential to lead to inconsistent and unfair allocations that may jeopardise the perceived legitimacy of the priority setting process. It was also argued that democratic or Communitarian approaches to estimating explicit equity weights may enhance societal well-being by aligning healthcare outcomes more closely with societal preferences.

The societal preferences estimated here were particularly strong over patient age and the quality of a patient's final health state, suggesting that societal well-being may be enhanced by giving priority to younger patients and those more likely to finish treatment in a good health state. This priority may be in the form of greater weight to these characteristics in an equity-weighted QALY, although such priority must also be consistent with ethical and legal

frameworks that ensure that the allocation of healthcare over these characteristics is just. It is also important to recognise that there were some substantial discrepancies between the two elicitation methods used here in the strength and significance of the equity-efficiency trade-off over different attributes. This inconsistency was mirrored by the heterogeneity in the results of the larger body of societal preferences research, and this made it difficult to identify a societally preferred weighting scheme. This consistent inconsistency, although perhaps reflective of some procedural variance in the estimation of preferences, may also suggest that a single set of societal weights may not exist, and in this light efforts at explicit equity-weighting on the basis of current research may represent a second-best solution that could worsen rather than improve the allocation of healthcare resources.

Some encouragement, though, may be drawn from Sen (1992), who noted that “an approach that can rank the well-being of every person against every other in a straightforward way... may well be at odds with the nature of these ideas.” In this sense, evidence of heterogeneity may not be so much a fatal flaw of explicit or participatory approaches to priority setting, but rather an inherent property of any method that seeks to incorporate individual preferences. The goal of research in this area, therefore, should be to contribute to what Ham and Coulter (2001) described as “the challenge of improving both technical and decision-making processes to enable the judgements that lie behind rationing to be as soundly based as possible.”

Appendix 10.1: Summary of recent DCE and CSPC stated preference elicitations

| Study | Setting | Method | Consistent w/ QALY max? | LYs/QALYs | Age | Life expectancy | Initial QoL | Final QoL | Responsibility for illness | SES | Patients treated | Previous healthcare | Other treatments | Disease prevalence | Social role | Time waited | Female | Chance of success | Value for money | Attributes | | | | |
|--|------------|--------|----------------------------|-----------|-----|-----------------|-------------|-----------|----------------------------|-----|------------------|---------------------|------------------|--------------------|-------------|-------------|--------|-------------------|-----------------|------------|--|--|---|-----|
| | | | | | | | | | | | | | | | | | | | | | | | | |
| Nord (1996) | Australia | PTO | | ↑ | ↑ | | | | | | | | | | | | | | | | | | | |
| Abellan-Perpinan & Pinto-Prades (1999) | Spain | CSPC | | | | | ns | | | | | | | | | | | | | | | | | |
| Ratcliffe (2000) | UK* | CSPC | | ↑ | ↓ | | | | ↓ | | | ↑ | | | | ↑ | | | | | | | | |
| Bryan et al. (2002) | UK | DCE | Yes | ↑ | | | ↑ | | | | ↑ | | | | | | | | | | | | ↑ | |
| Schwappach (2003) | Germany | CSPC | No | ↑ | ↓ | ↑ | ↑ | | ↑ | | | ↓ | | | | | | | | | | | | |
| Dolan & Tsuchiya (2005) | UK | DCE | | | ↓ | ↑ | ns | | | | | | | | | | | | | | | | | |
| Baltussen et al. (2006) | Ghana | DCE | Yes | | ↓ | | ↓ | | | | ↑ns | | | | | | | | | | | | | ↑ |
| Chan (2006) | Hong Kong* | CSPC | | ↑ | ↓ | | | | ↓ | | | ↓ | | | | ↑ | | | | | | | | |
| Dolan et al. (2008) | UK | DCE | No | | ↓ | | ↓ | | ↓ | | | | | | | | | | | | | | | ↓ns |
| Green & Gerard (2009) | UK | DCE | Yes | ↑ | | | ↓ | | | | | | ↓ns | | | | | | | | | | | ↑ |
| Baker et al. (2010) | UK | PTO | | | ↓ | | ∩ | | | | | | | | | | | | | | | | | |
| Desser et al. (2010) | Norway† | CSPC | | | | | | | | | ns | | | | | | | | | | | | | ns |

| Koopmanschap et al. (2010) | Netherlands | DCE | Yes | ↑ | ↓ | ↑ns | ↑ |
|----------------------------|-------------|------|-----|---|-----|-----|-----|
| Lancsar et al. (2011) | UK | DCE | Yes | ↑ | ↓ | ↑ns | ↑ns |
| Diederich et al. (2012) | Germany | DCE | | ∩ | ↓ | ns | ↓ |
| Linley & Hughes (2012) | UK† | CSPC | No | ↑ | ↓ns | ↓ns | ↓ |
| Norman et al. (2013) | Australia | DCE | No | ↑ | ↓ | ↓ | ↑ |
| Shah et al. (2012) | UK | DCE | | ↑ | ↑ | ↑ | ns |
| Skedgel | Canada‡ | DCE | | ↑ | ↓ | ↑ | ↑ns |
| Skedgel | Canada‡ | CSPC | | ↑ | ↓ | ns | ↓ |

Consistent w/ QALY max? = Authors' assessment of consistency of elicited preferences with principles of QALY maximisation; QoL=Quality-of-life; SES=Socioeconomic status; ↑ = Respondents preferred to prioritise higher levels of attribute; ↓ = Respondents preferred to prioritise lower levels of attribute; ∩ = Non-linear preference peaking at middle levels and declining at upper and lower levels; ns=Attribute not statistically significant. Preferred levels are shown for non-significant attributes that appeared to have a discernible trend in direction of preference.

Specific disease contexts:
* Kidney transplant
† Orphan diseases
‡ Cancer

References

- Abellan-Perpinan, J.M. & Pinto-Prades, J.L., 1999. Health state after treatment: a reason for discrimination? *Health Economics*, 8(8), pp.701–707.
- Abelson, J., Forest, P.-G., et al., 2003. Deliberations about deliberative methods: issues in the design and evaluation of public participation processes. *Social science & medicine*, 57(2), pp.239–251.
- Abelson, J., Eyles, J., et al., 2003. Does deliberation make a difference? Results from a citizens panel study of health goals priority setting. *Health Policy*, 66(1), pp.95–106.
- Abelson, J. et al., 1995. Does the community want devolved authority? Results of deliberative polling in Ontario. *CMAJ: Canadian Medical Association Journal*, 153(4), pp.403–412.
- Ben-Akiva, M. et al., 1997. Modeling Methods for Discrete Choice Analysis. *Marketing Letters*, 8(3), pp.273–286.
- Ben-Akiva, M., Morikawa, T. & Shiroishi, F., 1991. Analysis of the reliability of preference ranking data. *Journal of Business Research*, 23(3), pp.253–268.
- Alexander, L. & Moore, M., 2008. Deontological Ethics. In E. N. Zalta, ed. *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. Available at: <http://plato.stanford.edu/entries/ethics-deontological/> [Accessed September 29, 2010].
- Ali, S. & Ronaldson, S., 2012. Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. *British Medical Bulletin*, 103(1), pp.21–44.
- Amaya-Amaya, M., Gerard, K. & Ryan, M., 2008. Discrete Choice Experiments in a Nutshell. In M. Ryan, K. Gerard, & M. Amaya-Amaya, eds. *Using Discrete Choice Experiments to Value Health and Health Care*. The Economics of Non-Market Goods and Resources. Springer Netherlands, pp. 13–46.
- Anand, P. & Wailoo, A., 2000. Utilities versus Rights to Publicly Provided Goods: Arguments and Evidence from Health Care Rationing. *Economica*, 67(268), pp.543–577.
- Anand, S., 2002. The concern for equity in health. *Journal of Epidemiology and Community Health*, 56(7), pp.485–487.
- Anderson, E.S., 1999. What Is the Point of Equality? *Ethics*, 109(2), p.287.

- Anon, 1985. *Canadian Human Rights Act*, Available at: <http://laws-lois.justice.gc.ca/eng/acts/H-6/FullText.html> [Accessed November 27, 2013].
- Anon, 2010. *Equality Act 2010*, Available at: <http://www.legislation.gov.uk/ukpga/2010/15/contents> [Accessed November 27, 2013].
- Appleby, J. & Harrison, A., 2006. *Spending on Health Care: How much is enough?*, London: King's Fund.
- Arneson, R.J., 2000. Luck egalitarianism and prioritarianism. *Ethics*, 110(2), pp.339–349.
- Arrow, K. et al., 1993. Report of the NOAA panel on contingent valuation. *Federal register*, 58(10), pp.4601–4614.
- Arrow, K.J., 1963. Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*, 53(5), pp.941–973.
- Auger, P., Devinney, T. & Louviere, J., 2007. Using Best-Worst Scaling Methodology to Investigate Consumer Ethical Beliefs Across Countries. *Journal of Business Ethics*, 70(3), pp.299–326.
- Australian Government Department of Health and Ageing, 2008. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.3). Available at: <http://www.pbs.gov.au/industry/listing/elements/pbac-guidelines/PBAC4.3.2.pdf> [Accessed July 20, 2013].
- Baker, R. et al., 2010. Weighting and valuing quality-adjusted life-years using stated preference methods: preliminary results from the Social Value of a QALY Project. *Health Technology Assessment*, 14(27), pp.1–162.
- Baltagi, B.H., 2008. *Econometric Analysis of Panel Data* 4th ed., Chichester, UK: Wiley.
- Baltussen, R. et al., 2006. Towards a multi-criteria approach for priority setting: an application to Ghana. *Health Economics*, 15(7), pp.689–696.
- Baron, J. et al., 2001. Analog Scale, Magnitude Estimation, and Person Trade-off as Measures of Health Utility: Biases and their Correction. *Journal of Behavioral Decision Making*, 14(1), pp.17–34.
- Baron, J. & Greene, J., 1996. Determinants of insensitivity to quantity in valuation of public goods: Contribution, warm glow, budget constraints, availability, and prominence. *Journal of Experimental Psychology: Applied*, 2(2), pp.107–125.
- Bartels, D.M. & Medin, D.L., 2007. Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological science : a journal of the American Psychological Society / APS*, 18(1), pp.24–28.

- Beauchamp, T.L. & DeGrazia, D., 2004. Principles and Principlism. In G. Khushf, ed. *Handbook of Bioethics*. Springer, pp. 55–74.
- Bech, M. & Gyrd-Hansen, D., 2005. Effects coding in discrete choice experiments. *Health economics*, 14(10), pp.1079–1083.
- De Bekker-Grob, E.W., Hol, L., et al., 2010. Labeled versus Unlabeled Discrete Choice Experiments in Health Economics: An Application to Colorectal Cancer Screening. *Value in Health*, 13(2), pp.315–323.
- De Bekker-Grob, E.W., Ryan, M. & Gerard, K., 2010. Discrete choice experiments in health economics: a review of the literature. *Health Economics*, 21(2), pp.145–172.
- Bleichrodt, H. & Pinto, J.L., 2006. Conceptual foundations for health utility measurement. In A. M. Jones, ed. *The Elgar companion to health economics*. Cheltenham: Edward Elgar Publishing, p. 347.
- Borm, G.F. et al., 2005. Sequential balancing: a simple method for treatment allocation in clinical trials. *Contemporary clinical trials*, 26(6), pp.637–645.
- Bowling, A., 1996. Health care rationing: the public's debate. *BMJ*, 312(7032), pp.670–674.
- Boxall, P.C. & Adamowicz, W.L., 2002. Understanding Heterogeneous Preferences in Random Utility Models: A Latent Class Approach. *Environmental and Resource Economics*, 23(4), pp.421–446.
- Boyce, R.R. et al., 1992. An Experimental Examination of Intrinsic Values as a Source of the WTA-WTP Disparity. *The American Economic Review*, 82(5), pp.1366–1373.
- Boyle, K.J. et al., 2001. A Comparison of Conjoint Analysis Response Formats. *American Journal of Agricultural Economics*, 83(2), p.441.
- Boyle, K.J. & Ozdemir, S., 2009. Convergent Validity of Attribute-Based, Choice Questions in Stated-Preference Studies. *Environmental & Resource Economics*, 42(2), pp.247–264.
- Brandstatter, E., Gigerenzer, G. & Hertwig, R., 2006. The priority heuristic: making choices without trade-offs. *Psychological review*, 113(2), pp.409–432.
- Brazier, J., Deverill, M. & Green, C., 1999. A review of the use of health status measures in economic evaluation. *Journal of health services research & policy*, 4(3), pp.174–184.
- Bridges, J.F.P. et al., 2010. Conjoint Analysis Good Research Practices Task Force. Available at: <http://www.ispor.org/taskforces/conjointanalysisgrp.asp> [Accessed November 20, 2010].

- Broome, J., 1989. What's the point of equality? In J. D. Hey, ed. *Current issues in microeconomics*. New York: St. Martin's Press.
- Broqvist, M. & Garpenby, P., 2014. To accept, or not to accept, that is the question: citizen reactions to rationing. *Health Expectations*, 17(1), pp.82–92.
- Brouwer, W.B.F. et al., 2008. Welfarism vs. extra-welfarism. *Journal of Health Economics*, 27(2), pp.325–338.
- Brown, S., Harris, M.N. & Taylor, K., 2010. Modelling charitable donations: A latent class panel approach. Available at: <http://www.shef.ac.uk/economics/research/serps/year.html> [Accessed June 19, 2012].
- Bryan, S. et al., 2002. QALY-maximisation and public preferences: results from a general population survey. *Health economics*, 11(8), pp.679–693.
- Buchanan, J.T., Henig, E.J. & Henig, M.I., 1998. Objectivity and subjectivity in the decision making process. *Annals of Operations Research*, 80(0), pp.333–345.
- Burnham, K.P. & Anderson, D.R., 2004. Multimodel Inference Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2), pp.261–304.
- Busschbach, J.J., Helsing, D.J. & Charro, F.T. de, 1993. The utility of health at different stages in life: a quantitative approach. *Social science & medicine (1982)*, 37(2), pp.153–158.
- Buxton, M.J. & Chambers, J.D., 2011. What values do the public want their health care systems to use in evaluating technologies? *European Journal of Health Economics*, 12(4), pp.285–288.
- Cairns, J., Pol, M. van der & Lloyd, A., 2002. Decision making heuristics and the elicitation of preferences: being fast and frugal about the future. *Health Economics*, 11(7), pp.655–658.
- Callahan, D., 2003a. Individual Good and Common Good: A Communitarian Approach to Bioethics. *Perspectives in biology and medicine*, 46(4), p.496.
- Callahan, D., 2003b. Principlism and communitarianism. *Journal of medical ethics*, 29(5), pp.287–291.
- Canadian Agency For Drugs and Technologies in Health, 2013. *Common Drug Review Procedures*, Canadian Agency for Drugs and Technologies in Health. Available at: http://www.cadth.ca/media/cdr/process/CDR_Procedure_e.pdf [Accessed July 3, 2013].

- Canadian Agency For Drugs and Technologies in Health, 2006. *Guidelines for the Economic Evaluation of Health Technologies: Canada* 3rd ed., Ottawa: Canadian Agency for Drugs and Technology in Health.
- Carlsson, F. & Martinsson, P., 2003. Design techniques for stated preference methods in health economics. *Health Economics*, 12(4), pp.281–294.
- Carruthers, S.I. & Ormondroyd, J., 2009. Age equality in health and social care. Available at: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_107278 [Accessed November 27, 2013].
- Carson, R.T. et al., 1994. Experimental analysis of choice. *Marketing Letters*, 5(4), pp.351–367.
- Carson, R.T., Flores, N.E. & Meade, N.F., 2001. Contingent Valuation: Controversies and Evidence. *Environmental and Resource Economics*, 19(2), pp.173–210; 210.
- Carson, R.T. & Louviere, J.J., 2011. A Common Nomenclature for Stated Preference Elicitation Approaches. *Environmental and Resource Economics*, 49(4), pp.539–559.
- Chan, H.M., Cheung, G.M.Y. & Yip, A.K.W., 2006. Selection criteria for recipients of scarce donor livers: a public opinion survey in Hong Kong. *Hong Kong Medical Journal*, 12(1), pp.40–46.
- Chien, Y.-L., Huang, C.J. & Shaw, D., 2005. A general model of starting point bias in double-bounded dichotomous contingent valuation surveys. *Journal of Environmental Economics and Management*, 50(2), pp.362–377.
- Choudhry, N. et al., 1997. Distributional dilemmas in health policy: large benefits for a few or smaller benefits for many? *Journal of Health Services Research Policy*, 2(4), p.212.
- Clark, S. & Muthen, B., 2009. Relating latent class analysis results to variables not included in the analysis. Available at: <http://statmodel2.com/download/relatinglca.pdf> [Accessed September 27, 2012].
- Coast, J., 2001a. Citizens, their agents and health care rationing: an exploratory study using qualitative methods. *Health economics*, 10(2), pp.159–174.
- Coast, J., 2004. Is economic evaluation in touch with society's health values? *BMJ*, 329(7476), pp.1233–1236.
- Coast, J., 2009. Maximisation in extra-welfarism: A critique of the current position in health economics. *Social science & medicine*, 69(5), pp.786–792.
- Coast, J., 1997. The rationing debate. Rationing within the NHS should be explicit. The case against. *BMJ*, 314(7087), pp.1118–1122.

- Coast, J., 2001b. Who wants to know if their care is rationed? Views of citizens and service informants. *Health Expectations*, 4(4), pp.243–252.
- Coast, J. & Horrocks, S., 2007. Developing attributes and levels for discrete choice experiments using qualitative methods. *Journal of health services research & policy*, 12(1), pp.25–30.
- Coast, J., Smith, R. & Lorgelly, P., 2008a. Should the capability approach be applied in Health Economics? *Health Economics*, 17(6), pp.667–670.
- Coast, J., Smith, R. & Lorgelly, P., 2008b. Welfarism, extra-welfarism and capability: The spread of ideas in health economics. *Social science & medicine*, 67(7), pp.1190–1198.
- Cohen, J., 1988. *Statistical power analysis for the behavioral sciences*, Hillsdale, N.J.: L. Erlbaum Associates.
- Cookson, R., 2005. QALYs and the capability approach. *Health economics*, 14(8), p.817.
- Cookson, R. & Dolan, P., 2000. Principles of justice in health care rationing. *Journal of medical ethics*, 26(5), pp.323–329.
- Cookson, R. & Dolan, P., 1999. Public views on health care rationing: a group discussion study. *Health Policy*, 49(1-2), pp.63–74.
- Croissant, Y., 2012. *mlogit: multinomial logit model*, Available at: <http://CRAN.R-project.org/package=mlogit>.
- Croissant, Y. & Millo, G., 2008. Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*, 27(2), pp.1–43.
- Culyer, A.J., 2001a. Economics and ethics in health care. *Journal of medical ethics*, 27(4), pp.217–222.
- Culyer, A.J., 2001b. Equity - some theory and its policy implications. *Journal of medical ethics*, 27(4), pp.275–283.
- Culyer, A.J., 1989. The normative economics of health care finance and provision. *Oxford Review of Economic Policy*, 5(1), p.34.
- Culyer, A.J. & Wagstaff, A., 1993. Equity and equality in health and health care. *Journal of health economics*, 12(4), pp.431–457.
- D'Agostino, F. & Gaus, G., 2008. Contemporary Approaches to the Social Contract. In E. N. Zalta, ed. *The Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. Available at: <http://plato.stanford.edu/entries/contractarianism-contemporary/> [Accessed September 29, 2010].

- Damschroder, L.J. et al., 2004. The Validity of Person Tradeoff Measurements: Randomized Trial of Computer Elicitation Versus Face-to-Face Interview. *Medical Decision Making*, 24(2), pp.170–180.
- Damschroder, L.J. et al., 2005. Trading people versus trading time: what is the difference? *Population health metrics*, 3, p.10.
- Damschroder, L.J. et al., 2007. Why people refuse to make tradeoffs in person tradeoff elicitation: a matter of perspective? *Medical Decision Making*, 27(3), pp.266–280.
- Daniels, N., 1998. Distributive justice and the use of summary measures of population health status. In M. J. Field & M. R. Gold, eds. *Summarizing Population Health: Directions for the development and application of population metrics*. Washington, D.C.: National Academy Press, pp. 58–71.
- Daniels, N., 1990. Equality of What: Welfare, Resources, or Capabilities? *Philosophy and Phenomenological Research*, 50(Supplement), pp.273–296.
- Daniels, N., 2001. Justice, health, and healthcare. *The American journal of bioethics : AJOB*, 1(2), pp.2–16.
- Daniels, N. & Sabin, J.E., 2002. *Setting limits fairly: Can we learn to share medical resources?*, New York: Oxford University Press, USA.
- Desarbo, W. et al., 1997. Representing Heterogeneity in Consumer Response Models. *Marketing Letters*, 8(3), pp.335–348.
- DeShazo, J.R. & Fermo, G., 2002. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. *Journal of Environmental Economics and Management*, 44(1), p.123.
- Desser, A.S. et al., 2010. Societal views on orphan drugs: cross sectional survey of Norwegians aged 40 to 67. *BMJ*, 341(sep22 3), pp.c4715–c4715.
- Devlin, N., Appleby, J. & Parkin, D., 2003. Patients' views of explicit rationing: what are the implications for health service decision-making? *Journal of Health Services Research & Policy*, 8(3), pp.183–186.
- Dias, J.G. & Vermunt, J.K., 2006. Bootstrap methods for measuring classification uncertainty in latent class analysis. In A. Rizzi & M. Vichi, eds. *Compstat 2006 - Proceedings in Computational Statistics*. pp. 31–41. Available at: <http://www.springerlink.com/content/x17631r41t304v37/abstract/> [Accessed September 26, 2012].
- Diederich, A., Swait, J.D. & Wirsik, N., 2012. Citizen Participation in Patient Prioritization Policy Decisions: An Empirical and Experimental Study on Patients' Characteristics D. W. Dowdy, ed. *PLoS ONE*, 7(5), p.e36824.

- Dolan, P. et al., 2007. It ain't what you do, it's the way that you do it: Characteristics of procedural justice and their importance in social decision-making. *Journal of Economic Behavior & Organization*, 64(1), pp.157–170.
- Dolan, P., 2001. Output measures and valuation in health. In M. F. Drummond & A. McGuire, eds. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press.
- Dolan, P. et al., 2005. QALY maximisation and people's preferences: a methodological review of the literature. *Health Economics*, 14(2), pp.197–208.
- Dolan, P. et al., 2008. *The relative societal value of health gains to different beneficiaries*, Health Economics and Decision Science Discussion Paper Series. Available at: http://eprints.whiterose.ac.uk/11213/1/HEDS_DP_08-12.pdf [Accessed September 29, 2010].
- Dolan, P. & Cookson, R., 2000. A qualitative study of the extent to which health gain matters when choosing between groups of patients. *Health Policy*, 51(1), pp.19–30.
- 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(7188), pp.916–919.
- Dolan, P. & Green, C., 1998. Using the Person Trade-Off Approach to Examine Differences between Individual and Social Values. *Health economics*, 7(4), pp.307–312.
- Dolan, P. & Shaw, R., 2004. A note on a discussion group study of public preferences regarding priorities in the allocation of donor kidneys. *Health Policy*, 68(1), pp.31–36.
- Dolan, P. & Tsuchiya, A., 2005. Health priorities and public preferences: the relative importance of past health experience and future health prospects. *Journal of health economics*, 24(4), pp.703–714.
- Donaldson, C. & Shackley, P., 1997. Does “process utility” exist? A case study of willingness to pay for laparoscopic cholecystectomy. *Social Science & Medicine*, 44(5), pp.699–707.
- Dowie, J., 1998. Towards the equitably efficient and transparently decidable use of public funds in the deep blue millennium. *Health Economics*, 7(2), pp.93–103.
- Doyal, L., 1995. Needs, rights, and equity: more quality in healthcare rationing. *Quality in health care : QHC*, 4(4), pp.273–283.

- Doyal, L., 1997. The rationing debate. Rationing within the NHS should be explicit. The case for. *BMJ*, 314(7087), pp.1114–1118.
- Drummond, M. et al., 2003. Use of pharmacoeconomics information—report of the ISPOR Task Force on use of pharmacoeconomic/health economic information in health-care decision making. *Value in Health*, 6(4), pp.407–416.
- Dworkin, R., 1989. The Original Position. In N. Daniels, ed. *Reading Rawls: critical studies on Rawls' A theory of justice*. Stanford: Stanford University Press, pp. 16–53.
- Edelman, 2013. Executive Summary: 2013 Edelman Trust Barometer. *Scribd*. Available at: <http://www.scribd.com/doc/121501475/Executive-Summary-2013-Edelman-Trust-Barometer> [Accessed September 13, 2013].
- Edlin, R., Tsuchiya, A. & Dolan, P., 2012. Public Preferences for Responsibility Versus Public Preferences for Reducing Inequalities. *Health Economics*, 21(12), pp.1416–1426.
- Edwards, R.T. et al., 2003. Clinical and lay preferences for the explicit prioritisation of elective waiting lists: survey evidence from Wales. *Health Policy*, 63(3), pp.229–237.
- Feiring, E., 2008. Lifestyle, responsibility and justice. *Journal of medical ethics*, 34(1), pp.33–36.
- Feldman, A. & Serrano, R., 2006. *Welfare Economics and Social Choice Theory, 2nd Edition*, Springer.
- Fine, A., 1998. The Viewpoint of No-One in Particular. *Proceedings and Addresses of the American Philosophical Association*, 72(2), pp.7–20.
- Fischer, G.W. et al., 1999. Goal-Based Construction of Preferences: Task Goals and the Prominence Effect. *Management Science*, 45(8), pp.1057–1075.
- Fischhoff, B. et al., 1993. Embedding effects: Stimulus representation and response mode. *Journal of Risk and Uncertainty*, 6(3), pp.211–234.
- Fleck, L.M., 1992. Just Health Care Rationing: A Democratic Decisionmaking Approach. *University of Pennsylvania Law Review*, 140(5), pp.1597–1636.
- Flynn, T., 2010. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. *Expert review of pharmacoeconomics outcomes research*, 10(3), p.259.
- Flynn, T.N. et al., 2007. Best–worst scaling: What it can do for health care research and how to do it. *Journal of health economics*, 26(1), pp.171–189.
- Fox, J., 2003. Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15), pp.1–27.

- Fox, J. & Weisberg, S., 2011. *An R Companion to Applied Regression* Second., Thousand Oaks CA: Sage. Available at: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion> [Accessed January 29, 2013].
- Fraenkel, L., 2013. Incorporating Patients' Preferences Into Medical Decision Making. *Medical Care Research and Review*, 70(1 suppl), p.80S–93S.
- Freedman, D.A., 2006. On The So-Called “Huber Sandwich Estimator” and “Robust Standard Errors.” *The American Statistician*, 60(4), pp.299–302.
- Frey, B.S. & Stutzer, A., 2005. Beyond outcomes: measuring procedural utility. *Oxford Economic Papers*, 57(1), pp.90–111.
- Froberg, D.G. & Kane, R.L., 1989. Methodology for measuring health-state preferences–I: Measurement strategies. *Journal of clinical epidemiology*, 42(4), pp.345–354.
- Gafni, A., 1995. Time in health: can we measure individuals' “pure time preferences”? *Medical Decision Making*, 15(1), pp.31–37.
- Gericke, C.A., Riesberg, A. & Busse, R., 2005. Ethical issues in funding orphan drug research and development. *Journal of Medical Ethics*, 31(3), pp.164 – 168.
- Giacomini, M., Hurley, J. & DeJean, D., 2012. Fair reckoning: a qualitative investigation of responses to an economic health resource allocation survey. *Health Expectations*.
- Gigerenzer, G. & Gaissmaier, W., 2011. Heuristic Decision Making. *Annual Review of Psychology*, 62, pp.451–482.
- Gigerenzer, G. & Goldstein, D.G., 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological review*, 103(4), pp.650–669.
- Glasgow, G., 2001. Mixed Logit Models for Multiparty Elections. *Political Analysis*, 9(2), pp.116–136.
- Gold, M.R., 1996. *Cost-effectiveness in health and medicine*, New York: Oxford University Press.
- Goodin, R., 1986. Laundering preferences. In R. Elster & A. Hylland, eds. *Foundations of social choice theory*. New York: Cambridge University Press.
- Government of Canada, S.C., 2012. Population by sex and age group. Available at: <http://www.statcan.gc.ca/tables-tableaux/sum-som/101/cst01/demo10a-eng.htm> [Accessed August 2, 2012].
- Green, C., 2009. Investigating public preferences on “severity of health” as a relevant condition for setting healthcare priorities. *Social science & medicine* (1982), 68(12), pp.2247–2255.

- Green, C., 2001. On the societal value of health care: what do we know about the person trade-off technique? *Health economics*, 10(3), pp.233–243.
- Green, C. & Gerard, K., 2009. Exploring the social value of health-care interventions: a stated preference discrete choice experiment. *Health economics*, 18(8), pp.951–976.
- Greene, W.H. & Hensher, D.A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8), pp.681–698.
- Gregory, R., Lichtenstein, S. & Slovic, P., 1993. Valuing environmental resources: A constructive approach. *Journal of Risk and Uncertainty*, 7(2), pp.177–197.
- Gujarati, D.N., 1988. *Basic econometrics*, New York: McGraw-Hill.
- Guttman, N. et al., 2008. What should be given a priority – costly medications for relatively few people or inexpensive ones for many? The Health Parliament public consultation initiative in Israel. *Health Expectations*, 11(2), pp.177–188.
- Hall, J. et al., 2004. Using stated preference discrete choice modeling to evaluate health care programs. *Journal of Business Research*, 57(9), pp.1026–1032.
- Ham, C., 1998. Retracing the Oregon trail: the experience of rationing and the Oregon health plan. *BMJ: British Medical Journal*, 316(7149), pp.1965–1969.
- Ham, C. & Coulter, A., 2001. Explicit and implicit rationing: taking responsibility and avoiding blame for health care choices. *Journal of health services research & policy*, 6(3), pp.163–169.
- Harris, J., 2005. It's not NICE to discriminate. *Journal of medical ethics*, 31(7), pp.373–375.
- Harris, J., 1987. QALYfying the value of life. *Journal of medical ethics*, 13(3), pp.117–123.
- Harris, J., 1985. *The Value of Life: Introduction to Medical Ethics*, London: Routledge.
- Hasman, A., 2003. Eliciting reasons: Empirical methods in priority setting. *Health care analysis*, 11(1), pp.41–58.
- Hauck, K., Smith, P.C. & Goddard, M., 2004. *The economics of priority setting for health care: a literature review*, World Bank.
- Hausman, D.M., 2002. The limits to empirical ethics. In C. J. L. Murray, J. A. Salomon, & C. D. Mathers, eds. *Summary measures of population health: concepts, ethics, measurement and applications*. Geneva: WHO, pp. 663–668.

- Hausman, D.M. & McPherson, M.S., 2006. *Economic Analysis, Moral Philosophy and Public Policy* 2nd ed., Cambridge, UK: Cambridge University Press.
- Hausman, D.M. & McPherson, M.S., 2009. Preference satisfaction and welfare economics. *Economics and Philosophy*, 25(01), p.1.
- Heldke, L.M. & Kellert, S.H., 1995. Objectivity as Responsibility. *Metaphilosophy*, 26(4), pp.360–378.
- Henningsen, A., 2012. *censReg: Censored Regression (Tobit) Models*, Available at: <http://CRAN.R-project.org/package=censReg> [Accessed January 3, 2013].
- Hensher, D.A. & Collins, A., 2011. Interrogation of Responses to Stated Choice Experiments: Is there sense in what respondents tell us? *Journal of Choice Modelling*, 4(1), pp.62–89.
- Hensher, D.A., Greene, W.H. & Rose, J.M., 2005. *Applied choice analysis: a primer*, Cambridge: Cambridge University Press.
- Hernández Alava, M., Wailoo, A.J. & Ara, R., 2012. Tails from the Peak District: Adjusted Limited Dependent Variable Mixture Models of EQ-5D Questionnaire Health State Utility Values. *Value in Health*, 15(3), pp.550–561.
- Herve, A., 2007. The Kendall Rank Correlation Coefficient. In Salkind, Neil J. & Rasmussen, Kristin, eds. *The Encyclopedia of Measurement and Statistics*. Thousand Oaks, Calif.: SAGE Publications, p. 1416.
- Hoffmann, C. et al., 2002. Do health-care decision makers find economic evaluations useful? The findings of focus group research in UK health authorities. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 5(2), pp.71–78.
- Hogarth, R.M. & Karelaia, N., 2005. Simple Models for Multiattribute Choice with Many Alternatives: When It Does and Does Not Pay to Face Trade-offs with Binary Attributes. *Management Science*, 51(12), pp.1860–1872.
- Hole, A.R., 2008. Modelling heterogeneity in patient preferences for the attributes of a general practitioner appointment. *Journal of Health Economics*, 27(4), pp.1078–1094.
- Hommel, G., 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2), pp.383–386.
- Hosmer, D.W. & Lemeshow, S., 2000. *Applied logistic regression*, New York: Wiley.
- Huber, J., 2009. *What we have learned from 20 years of conjoint research?*, Sawtooth Software. Available at:

<http://www.sawtoothsoftware.com/download/techpap/whatlrnd.pdf>
[Accessed September 29, 2010].

- Huber, J. & Zwerina, K., 1996. The Importance of Utility Balance in Efficient Choice Designs. *Journal of Marketing Research*, 33(3), p.307.
- Hughes, D.A., Tunnage, B. & Yeo, S.T., 2005. Drugs for exceptionally rare diseases: do they deserve special status for funding? *QJM*, 98(11), pp.829–836.
- Hunter, D.J., 2001. Rationing Healthcare: The Appeal of Muddling Through Elegantly. *HealthcarePapers*, 2(2), pp.31–37.
- Hurley, J., 1998. Welfarism, Extra-Welfarism and Evaluative Economic Analysis in the Health Sector. In T. E. Getzen, M. L. Barer, & G. L. Stoddart, eds. *Health, Health Care and Health Economics: Perspectives on Distribution*. Chichester, NY: John Wiley & Sons, p. 373.
- Innvaer, S. et al., 2002. Health policy-makers' perceptions of their use of evidence: a systematic review. *Journal of health services research & policy*, 7(4), pp.239–244.
- Jackman, S., 2001. Compulsory Voting. In *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, pp. 16314–16318.
- Jedidi, K. & Zhang, Z.J., 2002. Augmenting Conjoint Analysis to Estimate Consumer Reservation Price. *Management Science*, 48(10), pp.1350–1368.
- Johannesson, M. & Johansson, P.O., 1997. A note on prevention versus cure. *Health Policy*, 41(3), pp.181–187.
- Johansson-Stenman, O., 1998. On the problematic link between fundamental ethics and economic policy recommendations. *Journal of Economic Methodology*, 5(2), pp.263–297.
- Johnson, F. et al., 2007. Experimental Design For Stated-Choice Studies. In B. J. Kanninen, ed. *Valuing Environmental Amenities Using Stated Choice Studies*. The Economics of Non-Market Goods and Resources. Springer Netherlands, pp. 159–202.
- Kahneman, D. & Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), p.263.
- Kanninen, B.J., 2002. Optimal Design for Multinomial Choice Experiments. *Journal of Marketing Research*, 39(2), pp.214–227.
- Kaplan, R.M., Feeny, D. & Revicki, D.A., 1993. Methods for assessing relative importance in preference based outcome measures. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*, 2(6), pp.467–475.

- Kjær, T., 2005. *A review of the discrete choice experiment - with emphasis on its application in health care*, University of Southern Denmark. Available at: http://static.sdu.dk/mediafiles//Files/Om_SDU/Centre/c_ist_sundoke/Forskningsdokumenter/publications/Working papers/20051pdf.pdf [Accessed May 21, 2013].
- Klein, R., 1992. Dilemmas and decisions. *Health management quarterly: HMQ*, 14(2), pp.2–5.
- Klein, R., 1997. The rationing debate. Defining a package in healthcare services the NHS is responsible for. The case against. *BMJ*, 314(7079), pp.506–509.
- Klein, R. & Williams, A., 2000. Setting priorities: what is holding us back— inadequate information or inadequate institutions. In C. Ham & A. Coulter, eds. *The global challenge of health care rationing*. Buckingham: Open University Press, pp. 15–26.
- Klose, T., 1999. The contingent valuation method in health care. *Health policy*, 47(2), p.97.
- Konow, J., 2003. Which is the fairest one of all? A positive analysis of justice theories. *Journal of Economic Literature*, 41(4), pp.1188–1239.
- Koopmanschap, M.A., Stolk, E.A. & Koolman, X., 2010. Dear policy maker: Have you made up your mind? A discrete choice experiment among policy makers and other health professionals. *International Journal of Technology Assessment in Health Care*, 26(02), p.198.
- Kreuter, F., Presser, S. & Tourangeau, R., 2008. Social Desirability Bias in CATI, IVR, and Web Surveys. *Public Opinion Quarterly*, 72(5), pp.847 – 865.
- Kuhfeld, W.F., 2010. Marketing Research Methods in SAS. *Marketing Research Methods in SAS*. Available at: http://support.sas.com/resources/papers/tnote/tnote_marketresearch.html [Accessed December 1, 2010].
- Kuhfeld, W.F., Tobias, R.D. & Garratt, M., 1994. Efficient Experimental Design with Marketing Research Applications. *Journal of Marketing Research*, 31(4), pp.545–557.
- Lamont, J. & Favor, C., 2008. Distributive Justice. In E. N. Zalta, ed. *The Stanford Encyclopedia of Philosophy*. Available at: <http://plato.stanford.edu/archives/fall2008/entries/justice-distributive/> [Accessed January 31, 2012].
- Lancaster, K.J., 1966. A New Approach to Consumer Theory. *The Journal of Political Economy*, 74(2), pp.132–157.

- Lancsar, E. et al., 2011. Deriving distributional weights for QALYs through discrete choice experiments. *Journal of Health Economics*, 30(2), pp.466–478.
- Lancsar, E. & Louviere, J., 2006. Deleting “irrational” responses from discrete choice experiments: a case of investigating or imposing preferences? *Health Economics*, 15(8), pp.797–811.
- Lancsar, E., Louviere, J. & Flynn, T., 2007. Several methods to investigate relative attribute impact in stated preference experiments. *Social Science & Medicine*, 64(8), pp.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(9), pp.901–907.
- Lauridsen, S.M., Norup, M.S. & Rossel, P.J., 2007. The secret art of managing healthcare expenses: investigating implicit rationing and autonomy in public healthcare systems. *Journal of medical ethics*, 33(12), pp.704–707.
- Lee, J.A., Soutar, G.N. & Louviere, J., 2007. Measuring values using best-worst scaling: The LOV example. *Psychology and Marketing*, 24(12), pp.1043–1058.
- Leggett, C.G. et al., 2003. Social Desirability Bias in Contingent Valuation Surveys Administered Through In-Person Interviews. *Land Economics*, 79(4), pp.561–575.
- LeGrand, J., 1987. Equity, health, and health care. *Social Justice Research*, 1(3), pp.257–274.
- Leonard, E.W., 2012. *Death Panels and the Rhetoric of Rationing*, Rochester, NY: Social Science Research Network. Available at: <http://papers.ssrn.com/abstract=2147468> [Accessed September 13, 2013].
- Lim, J.N. & Edlin, R., 2009. Preferences of older patients and choice of treatment location in the UK: A binary choice experiment. *Health Policy*, 91(3), pp.252–257.
- Lindholm, L., Rosen, M. & Emmelin, M., 1998. How many lives is equity worth? A proposal for equity adjusted years of life saved. *Journal of epidemiology and community health*, 52(12), pp.808–811.
- Linley, W.G. & Hughes, D.A., 2012. Societal views on NICE, cancer drugs fund and value-based pricing criteria for prioritising medicines: a cross-sectional survey of 41 18 adults in Great Britain. *Health Economics*.
- List, J.A. & Gallet, C.A., 2001. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? *Environmental and Resource Economics*, 20(3), pp.241–254.

- Litva, A. et al., 2002. “The public is too subjective”: public involvement at different levels of health-care decision making. *Social Science & Medicine*, 54(12), pp.1825–1837.
- Lloyd, A.J., 2003. Threats to the estimation of benefit: are preference elicitation methods accurate? *Health Economics*, 12(5), pp.393–402.
- Lomas, J., 1997. Reluctant rationers: public input to health care priorities. *Journal of Health Services Research & Policy*, 2(2), pp.103–111.
- Long, J., 1997. *Regression models for categorical and limited dependent variables*, Thousand Oaks: Sage Publications.
- Loomis, J., 2011. What’s to know about hypothetical bias in stated preference valuation studies? *Journal of Economic Surveys*, 25(2), pp.363–370.
- Louviere, J.J., Hensher, D.A. & Swait, J.D., 2000a. Conjoint Preference Elicitation Methods in the Broader Context of Random Utility Theory Preference Elicitation Methods. In A. Gustafsson, A. Herrmann, & F. Huber, eds. *Conjoint measurement: methods and applications*. Berlin: Springer, pp. 167–198.
- Louviere, J.J., Hensher, D.A. & Swait, J.D., 2000b. *Stated choice methods: analysis and applications*, Cambridge, U.K.: Cambridge University Press.
- Louviere, J.J. & Islam, T., 2008. A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best-worst scaling. *Journal of Business Research*, 61(9), pp.903–911.
- Magidson, J. & Vermunt, J.K., 2004. Latent class models. In D. Kaplan, ed. *The Sage Handbook of Quantitative Methodology for the Social Sciences*. SAGE.
- Marley, A.A.J. & Louviere, J.J., 2005. Some probabilistic models of best, worst, and best-worst choices. *Journal of mathematical psychology*, 49(6), pp.464–480.
- Marshall, D.D. et al., 2010. Conjoint Analysis Applications in Health — How are Studies being Designed and Reported? *The Patient: Patient-Centered Outcomes Research*, 3(4), pp.249–256.
- Mathews, K.E., Freeman, M.L. & Desvousges, W.H., 2007. How and How Much? In *Valuing Environmental Amenities Using Stated Choice Studies*. The Economics of Non-Market Goods and Resources. Springer Netherlands, pp. 111–133.
- Mauskopf, J.A. et al., 1998. The role of cost-consequence analysis in healthcare decision-making. *Pharmacoeconomics*, 13(3), pp.277–288.
- Maynard, A. & Bloor, K., 1998. *Our Certain Fate: Rationing in Health Care*, London: Office of Health Economics.

- McCabe, C., 2006. Orphan drugs revisited. *QJM*, 99(5), pp.341–345.
- McFadden, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka, ed. *Frontiers in Econometrics*. New York: Academic Press, p. 252.
- McFadden, D., 1999. Rationality for Economists? *Journal of Risk and Uncertainty*, 19(1), pp.73–105.
- McIntosh, E., 2003. *Using discrete choice experiments to value the benefits of health care*. PhD Thesis. Aberdeen: University of Aberdeen.
- McIntosh, E. & Ryan, M., 2002. Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: Implications of discontinuous preferences. *Journal of Economic Psychology*, 23(3), pp.367–382.
- McKie, J. & Richardson, J., 2003. The rule of rescue. *Social Science & Medicine*, 56(12), pp.2407–2419.
- McQuillin, B. & Sugden, R., 2012. Reconciling normative and behavioural economics: the problems to be solved. *Social Choice and Welfare*, 38(4), pp.553–567.
- Mechanic, D., 1995. Dilemmas in rationing health care services: the case for implicit rationing. *BMJ*, 310(6995), pp.1655–1659.
- Mentzakis, E., Stefanowska, P. & Hurley, J., 2011. A Discrete Choice Experiment Investigating Preferences for Funding Drugs Used to Treat Orphan Diseases: An Exploratory Study. *Health Economics, Policy and Law*, 6(03), pp.405–433.
- Menzel, P., 1999. How should what economists call “social values” be measured. *The Journal of Ethics*, 3(3), p.249.
- Menzel, P. et al., 1999. Toward a broader view of values in cost-effectiveness analysis of health. *The Hastings Center report*, 29(3), pp.7–15.
- Mertz, M. et al., 2014. Research across the disciplines: a road map for quality criteria in empirical ethics research. *BMC Medical Ethics*, 15(1), p.17.
- Miethe, T.D., 1985. The Validity and Reliability of Value Measurements. *The Journal of Psychology*, 119(5), pp.441–453.
- Miguel, F.S., Ryan, M. & Amaya-Amaya, M., 2005. “Irrational” stated preferences: a quantitative and qualitative investigation. *Health Economics*, 14(3), pp.307–322.
- Mill, J.S., 1871. *Utilitarianism*, London: Longmans, Green, Reader, and Dyer.

- Miller, G.A., 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2), pp.81–97.
- Mitton, C.R., 2002. Priority setting for decision makers: using health economics in practice. *The European Journal of Health Economics*, 3(4), pp.240–243.
- Miyamoto, J.M. et al., 1998. The Zero-Condition: A Simplifying Assumption in QALY Measurement and Multiattribute Utility. *Management Science*, 44(6), pp.839–849.
- Miyamoto, J.M. & Eraker, S.A., 1988. A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology: General*, 117(1), pp.3–20.
- Mooney, G., 1998a. Beyond health outcomes: The benefits of health care. *Health Care Analysis*, 6(2), pp.99–105.
- Mooney, G., 2005. Communitarian claims and community capabilities: furthering priority setting? *Social Science & Medicine*, 60(2), pp.247–255.
- Mooney, G., 1998b. “Communitarian claims” as an ethical basis for allocating health care resources. *Social Science & Medicine*, 47(9), pp.1171–1180.
- Mooney, G., Jan, S. & Wiseman, V., 1995. Examining preferences for allocating health care gains. *Health Care Analysis*, 3(3), pp.261–265.
- Mooney, G. & Lange, M., 1993. Ante-natal screening: What constitutes “benefit”? *Social Science & Medicine*, 37(7), pp.873–878.
- Morey, E. & Greer Rossmann, K., 2003. Using Stated-Preference Questions to Investigate Variations in Willingness to Pay for Preserving Marble Monuments: Classic Heterogeneity, Random Parameters, and Mixture Models. *Journal of Cultural Economics*, 27(3), pp.215–229.
- Mortimer, D. & Segal, L., 2008. Is the value of a life or life-year saved context specific? Further evidence from a discrete choice experiment. *Cost Effectiveness and Resource Allocation*, 6(1), p.8.
- Mossialos, E. & King, D., 1999. Citizens and rationing: analysis of a European survey. *Health Policy*, 49(1-2), pp.75–135.
- Mueller, D.C., 2003. *Public choice III*, New York: Cambridge University Press.
- Mullen, P. & Spurgeon, P., 1999. *Priority Setting & The Public*, Abingdon, UK: Radcliffe Publishing Ltd.
- Mullen, P.M., 1999. Public involvement in health care priority setting: an overview of methods for eliciting values. *Health Expectations*, 2(4), pp.222–234.

- Murphy, J.J. et al., 2005. A Meta-analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics*, 30(3), pp.313–325.
- Murray, C.J.L. & Acharya, A.K., 1997. Understanding DALYs. *Journal of health economics*, 16(6), pp.703–730.
- National Institute for Health and Care Excellence, 2013. Citizens Council. Available at: <http://www.nice.org.uk/> [Accessed November 15, 2013].
- National Institute for Health and Care Excellence, 2008. *Citizens Council report on departing from the threshold*, Available at: <http://www.nice.org.uk/media/231/CB/NICECitizensCouncilDepartingThresholdFinal.pdf> [Accessed July 30, 2013].
- National Institute for Health and Clinical Excellence, 2009. Appraising life-extending, end of life treatments. Available at: <http://www.nice.org.uk/media/88A/F2/SupplementaryAdviceTACEoL.pdf> [Accessed August 16, 2013].
- National Institute for Health and Clinical Excellence, 2004. NICE Citizens Council Report: Ultra Orphan Drugs. Available at: http://www.nice.org.uk/niceMedia/pdf/Citizens_Council_Ultraorphan.pdf [Accessed August 17, 2013].
- National Institute for Health and Clinical Excellence, 2008. *Social value judgements: principles for the development of NICE guidance* 2nd ed., London: NICE.
- New, B., 2000. Commentary: an open debate is not an admission of failure. *BMJ*, 321(7252), p.45.
- Nord, E., 1996. Health status index models for use in resource allocation decisions. A critical review in the light of observed preferences for social choice. *International Journal of Technology Assessment in Health Care*, 12(1), pp.31–44.
- Nord, E. et al., 1999. Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics*, 8(1), pp.25–39.
- Nord, E. et al., 1995. Maximizing health benefits vs egalitarianism: An Australian survey of health issues. *Social Science & Medicine*, 41(10), pp.1429–1437.
- Nord, E., 1995a. The Person-Trade-Off Approach to Valuing Health-Care Programs. *Medical Decision Making*, 15(3), pp.201–208.
- Nord, E. et al., 1996. The significance of age and duration of effect in social evaluation of health care. *Health Care Analysis*, 4(2), pp.103–111.
- Nord, E., 1995b. The use of cost-value analysis to judge patients' right to treatment. *Medicine and law*, 14(7-8), pp.553–558.

- Norman, R. et al., 2013. Efficiency and Equity: A Stated Preference Approach. *Health Economics*, 22(5), pp.568–581.
- Normand, C., 2009. Measuring Outcomes in Palliative Care: Limitations of QALYs and the Road to PaLYs. *Journal of Pain and Symptom Management*, 38(1), pp.27–31.
- Nozick, R., 1974. *Anarchy, state, and utopia*, New York: Basic Books.
- Nussbaum, M.C., 2011. *Creating capabilities: the human development approach*, Cambridge, Mass: Belknap Press of Harvard University Press.
- Oehlert, G.W., 1992. A Note on the Delta Method. *The American Statistician*, 46(1), pp.27–29.
- Oliver, A., 2013. Testing Procedural Invariance in the Context of Health. *Health Economics*, 22(3), pp.272–288.
- Oliver, A., Mossialos, E. & Robinson, R., 2004. Health technology assessment and its influence on health-care priority setting. *International Journal of Technology Assessment in Health Care*, 20(1), pp.1–10.
- Olsen, J.A., 2000. A note on eliciting distributive preferences for health. *Journal of Health Economics*, 19(4), pp.541–550.
- Olsen, J.A. et al., 2003. The moral relevance of personal characteristics in setting health care priorities. *Social Science & Medicine*, 57(7), pp.1163–1172.
- Olsen, J.A., Richardson, J. & Mortimer, D., 1998. *Priority setting in the public health service: Results of Australian survey*, Melbourne: Centre for Health Program Evaluation, Monash University.
- Orme, B., 2006a. *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research*, Madison: Research Publishers LLC.
- Orme, B., 2006b. Sample Size Issues for Conjoint Analysis. In *Getting started with conjoint analysis : strategies for product design and pricing research*. Madison WI: Research Publishers LLC, pp. 57–66.
- Owen-Smith, A., Coast, J. & Donovan, J., 2010. The desirability of being open about health care rationing decisions: findings from a qualitative study of patients and clinical professionals. *Journal of Health Services Research & Policy*, 15(1), pp.14–20.
- pan-Canadian Oncology Drug Review, 2011. *pCODR Procedures*, Ottawa, Canada: Pan-Canadian Oncology Drug Review. Available at: <http://www.pcodr.ca/idc/groups/pcodr/documents/pcodrdocument/pcodr-procedures.pdf> [Accessed July 3, 2013].
- Parfit, D., 1997. Equality and Priority. *Ratio*, 10(3), pp.202–221.

- Payne, J.W. et al., 1992. A constructive process view of decision making: Multiple strategies in judgment and choice. *Acta Psychologica*, 80(1–3), pp.107–141.
- Payne, J.W., Bettman, J.R. & Johnson, E.J., 1993. *The adaptive decision maker*, New York, NY, USA: Cambridge University Press.
- Persad, G., Wertheimer, A. & Emanuel, E.J., 2009. Principles for allocation of scarce medical interventions. *The Lancet*, 373(9661), pp.423–431.
- Petrou, S. et al., 2013. A Person Trade-Off Study to Estimate Age-Related Weights for Health Gains in Economic Evaluation. *Pharmacoeconomics*, pp.1–15.
- Pinto Prades, J.L., 1997. Is the person trade-off a valid method for allocating health care resources? *Health Economics*, 6(1), pp.71–81.
- Pliskin, J.S., Shepard, D.S. & Weinstein, M.C., 1980. Utility Functions for Life Years and Health Status. *Operations Research*, 28(1), pp.206–224.
- Potoglou, D. et al., 2011. Best-worst scaling vs. discrete choice experiments: An empirical comparison using social care data. *Social Science & Medicine*, 72(10), pp.1717–1727.
- Price, D., 2000. Choices without reasons: citizens' juries and policy evaluation. *Journal of Medical Ethics*, 26(4), pp.272–276.
- Provencher, B., Baerenklau, K.A. & Bishop, R.C., 2002. A Finite Mixture Logit Model of Recreational Angling with Serially Correlated Random Utility. *American Journal of Agricultural Economics*, 84(4), pp.1066–1075.
- R Core Team, 2013. *R: A Language and Environment for Statistical Computing*, Vienna, Austria. Available at: <http://www.R-project.org/> [Accessed March 3, 2014].
- Ramaswamy, V. & Cohen, S.H., 2007. Latent class models for conjoint analysis. In A. Gustafsson, A. Herrmann, & F. Huber, eds. *Conjoint measurement methods and applications*. Berlin; New York: Springer. Available at: <http://dx.doi.org/10.1007/978-3-540-71404-0> [Accessed June 12, 2012].
- Ratcliffe, J. et al., 2009. Examining the attitudes and preferences of health care decision-makers in relation to access, equity and cost-effectiveness: a discrete choice experiment. *Health Policy*, 90(1), pp.45–57.
- Ratcliffe, J., 2000. Public preferences for the allocation of donor liver grafts for transplantation. *Health Economics*, 9(2), pp.137–148.
- Rawlins, M.D., 2005. Pharmacopolitics and deliberative democracy. *Clinical medicine*, 5(5), pp.471–475.
- Rawls, J., 1999. *A Theory of Justice*, Cambridge, Mass.: Belknap Press of Harvard University Press.

- Rawls, J., 2001. *Justice as Fairness: A Restatement*, Harvard University Press.
- Richardson, J., 1994. Cost utility analysis: what should be measured? *Social science & medicine* (1982), 39(1), pp.7–21.
- Richardson, J., 2002. The poverty of ethical analyses in economics and the unwarranted disregard of evidence. In C. J. L. Murray, J. A. Salomon, & C. D. Mathers, eds. *Summary measures of population health: concepts, ethics, measurement and applications*. Geneva: WHO, pp. 627–640.
- 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 & Medicine*, 60(2), pp.265–275.
- Rizopoulos, D., 2006. ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, 17(5), pp.1–25.
- Roberts, T. et al., 1999. Public involvement in health care priority setting: an economic perspective. *Health Expectations*, 2(4), pp.235–244.
- Robinson, R., 1999. Limits to rationality: economics, economists and priority setting. *Health Policy*, 49(1-2), pp.13–26.
- Rodriguez-Miguez, E. & Pinto-Prades, J.L., 2002. Measuring the social importance of concentration or dispersion of individual health benefits. *Health Economics*, 11(1), pp.43–53.
- Rosenhead, J., 1980. Planning Under Uncertainty: 1. The Inflexibility of Methodologies. *The Journal of the Operational Research Society*, 31(3), pp.209–216.
- Ross, J., 1995. The use of economic evaluation in health care: Australian decision makers' perceptions. *Health Policy*, 31(2), pp.103–110.
- Rumbold, B., Alakeson, V. & Smith, P.C., 2012. *Rationing health care: Is it time to set out more clearly what is funded by the NHS?*, Nuffield Trust. Available at: http://www.nuffieldtrust.org.uk/sites/files/nuffield/publication/rationing_health_care_240212.pdf [Accessed July 8, 2013].
- Ryan, M., 1994. Agency in Health Care: Lessons for Economists from Sociologists. *American Journal of Economics and Sociology*, 53(2), pp.207–217.
- Ryan, M., 2004. Deriving welfare measures in discrete choice experiments: a comment to Lancsar and Savage (1). *Health Economics*, 13(9), pp.909–12; discussion 919–24.
- Ryan, M. et al., 2001. Eliciting public preferences for healthcare: a systematic review of techniques. *Health technology assessment*, 5(5), pp.1–186.

- Ryan, M., 2009. Rationalising the “irrational”: A think aloud study of discrete choice experiment responses. *Health Economics*, 18(3), p.321.
- Ryan, M., 1999. Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. *Social Science & Medicine*, 48(4), pp.535–546.
- Ryan, M. & Gerard, K., 2003. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied Health Economics and Health Policy*, 2(1), pp.55–64.
- Ryan, M. & Skatun, D., 2004. Modelling non-demanders in choice experiments. *Health Economics*, 13(4), pp.397–402.
- Salkeld, G., 1998. What are the benefits of preventive health care? *Health Care Analysis*, 6(2), pp.106–112.
- Sassi, F., Archard, L. & LeGrand, J., 2001. Equity and the economic evaluation of healthcare. *Health Technology Assessment*, 5(3).
- Schenker, N. & Gentleman, J.F., 2001. On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. *The American Statistician*, 55(3), pp.182–186.
- Schlosser, R.W. et al., 2006. Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review. *International Journal of Language & Communication Disorders*, 41(5), pp.567–582.
- Schwappach, D.L., 2003. Does it matter who you are or what you gain? An experimental study of preferences for resource allocation. *Health Economics*, 12(4), pp.255–267.
- Schwappach, D.L., 2002a. Resource allocation, social values and the QALY: a review of the debate and empirical evidence. *Health Expectations*, 5(3), pp.210–222.
- Schwappach, D.L., 2002b. The equivalence of numbers: the social value of avoiding health decline: an experimental Web-based study. *BMC medical informatics and decision making*, 2, p.3.
- Schwappach, D.L. & Strasmann, T.J., 2006. “Quick and dirty numbers”? The reliability of a stated-preference technique for the measurement of preferences for resource allocation. *Journal of health economics*, 25(3), pp.432–448.
- Schwarzinger, M. et al., 2004. Lack of multiplicative transitivity in person trade-off responses. *Health Economics*, 13(2), pp.171–181.

- Scott, A., 2002. Identifying and analysing dominant preferences in discrete choice experiments: An application in health care. *Journal of Economic Psychology*, 23(3), pp.383–398.
- Sen, A., 2002. Why health equity? *Health economics*, 11(8), pp.659–666.
- Sen, A.K., 1985. *Commodities and capabilities*, Amsterdam: North-Holland.
- Sen, A.K., 1992. *Inequality reexamined*, New York: Oxford University Press.
- Sen, A.K., 2011. *The idea of justice*, Cambridge, Mass: Belknap Press of Harvard Univ. Press.
- Severin, V., 2001. *Comparing statistical and respondent efficiency in choice experiments*. Unpublished PhD dissertation. Sydney, Australia: University of Sydney.
- Shackley, P. & Ryan, M., 1995. Involving consumers in health care decision making. *Health Care Analysis*, 3(3), pp.196–204.
- Shaffer, J.P., 1995. Multiple Hypothesis Testing. *Annual Review of Psychology*, 46(1), pp.561–584.
- Shah, K. et al., 2012. *Valuing health at the end of life: a stated preference discrete choice experiment*, NICE Decision Support Unit. Available at: http://www.nicedsu.org.uk/DSU%20End%20of%20Life%20full%20report%20-%20version%203%20_Dec%202012_.pdf [Accessed December 6, 2012].
- Shah, K.K., 2009. Severity of illness and priority setting in healthcare: A review of the literature. *Health Policy*, 93(2-3), pp.77–84.
- Shen, J., 2009. Latent class model or mixed logit model? A comparison by transport mode choice data. *Applied Economics*, 41(22), pp.2915–2924.
- Shickle, D., 1997. Public Preferences for Health Care: Prioritisation in the United Kingdom. *Bioethics*, 11(3-4), pp.277–290.
- Shogren, J.F. et al., 1994. Resolving Differences in Willingness to Pay and Willingness to Accept. *The American Economic Review*, 84(1), pp.255–270.
- Silva, J.M.S., 2004. Deriving welfare measures in discrete choice experiments: a comment to Lancsar and Savage (2). *Health Economics*, 13(9), pp.913–8; discussion 919–24.
- Simon, H.A., 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1), pp.99–118.
- Slovic, P., 1995. The construction of preference. *American Psychologist*, 50(5), pp.364–371.
- Small, K.A. & Rosen, H.S., 1981. Applied Welfare Economics with Discrete Choice Models. *Econometrica*, 49(1), pp.105–130.

- Smith, R.D., 2000. The discrete-choice willingness-to-pay question format in health economics: Should we adopt environmental guidelines? *Medical Decision Making*, 20(2), pp.194–206.
- Stafinski, T. et al., 2011. Societal Values in the Allocation of Healthcare Resources: Is it All About the Health Gain? *The Patient*, 4(4), pp.207–225.
- Statistics Canada, 2010. Median total income, by family type, by province and territory. Available at: <http://www.statcan.gc.ca/tables-tableaux/sum-som/101/cst01/famil108a-eng.htm> [Accessed February 4, 2011].
- Statistics Canada, 2009. Population 15 years and over by highest degree, certificate or diploma (1986 to 2006 Census). Available at: <http://www.statcan.gc.ca/tables-tableaux/sum-som/101/cst01/educ42-eng.htm> [Accessed August 2, 2012].
- Sugden, R., 1993. Welfare, Resources, and Capabilities: A Review of Inequality Reexamined. *Journal of Economic Literature*, 31(4), pp.1947–1962.
- Sugden, R. & Williams, A.H., 1978. *The principles of practical cost-benefit analysis*, Oxford: Oxford University Press.
- Swait, J.D., 2007. Advanced Choice Models. In B. J. Kanninen, ed. *Valuing Environmental Amenities Using Stated Choice Studies*. The Economics of Non-Market Goods and Resources. Springer Netherlands, pp. 229–293.
- Swallow, S.K., Opaluch, J.J. & Weaver, T.F., 2001. Strength-of-preference indicators and an ordered-response model for ordinarily dichotomous, discrete choice data. *Journal of Environmental Economics and Management*, 41(1), pp.70–93.
- Tappenden, P. et al., 2007. A stated preference binary choice experiment to explore NICE decision making. *PharmacoEconomics*, 25(8), pp.685–693.
- Tsuchiya, A., 1999. Age-related preferences and age weighting health benefits. *Social Science & Medicine*, 48(2), pp.267–276.
- Tsuchiya, A., 2000. QALYs and ageism: philosophical theories and age weighting. *Health Economics*, 9(1), pp.57–68.
- Tsuchiya, A. & Dolan, P., 2009. Equality of what in health? Distinguishing between outcome egalitarianism and gain egalitarianism. *Health Economics*, 18(2), p.147.
- Tsuchiya, A. & Dolan, P., 2005. The QALY model and individual preferences for health states and health profiles over time: a systematic review of the literature. *Medical Decision Making*, 25(4), pp.460–467.
- Tsuchiya, A., Dolan, P. & Shaw, R., 2003. Measuring people's preferences regarding ageism in health: some methodological issues and some fresh evidence. *Social Science & Medicine*, 57(4), pp.687–696.

- Tsuchiya, A. & Williams, A., 2001. Welfare economics and economic evaluation. In M. F. Drummond & A. McGuire, eds. *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press.
- Tversky, A. & Kahneman, D., 1986. Rational Choice and the Framing of Decisions. *The Journal of Business*, 59(4, Part 2: The Behavioral Foundations of Economic Theory), pp.S251–S278.
- Ubel, P.A., DeKay, M.L., et al., 1996. Cost-effectiveness analysis in a setting of budget constraints—is it equitable? *The New England journal of medicine*, 334(18), pp.1174–1177.
- Ubel, P.A., Loewenstein, G., et al., 1996. Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon’s cost-effectiveness list failed. *Medical Decision Making*, 16(2), pp.108–116.
- Ubel, P.A. et al., 1998. Public preferences for prevention versus cure: what if an ounce of prevention is worth only an ounce of cure? *Medical Decision Making*, 18(2), pp.141–148.
- Ubel, P.A., Arnold, R.M. & Caplan, A.L., 1993. Rationing Failure: The Ethical Lessons of the Retransplantation of Scarce Vital Organs. *JAMA: The Journal of the American Medical Association*, 270(20), pp.2469–2474.
- Ubel, P.A., Baron, J. & Asch, D.A., 1999. Social Acceptability, Personal Responsibility, and Prognosis in Public Judgments and Transplant Allocation. *Bioethics*, 13(1), pp.57–68.
- Ubel, P.A. & Loewenstein, G., 1996. Distributing scarce livers: the moral reasoning of the general public. *Social science & medicine (1982)*, 42(7), pp.1049–1055.
- Ubel, P.A., Richardson, J. & Menzel, P., 2000. Societal value, the person trade-off, and the dilemma of whose values to measure for cost-effectiveness analysis. *Health Economics*, 9(2), pp.127–136.
- Ubel, P.A., Richardson, J. & Pinto-Prades, J.L., 1999. Life-saving treatments and disabilities. Are all QALYs created equal? *International Journal of Technology Assessment in Health Care*, 15(4), pp.738–748.
- Wagstaff, A., 1991. QALYs and the equity-efficiency trade-off. *Journal of health economics*, 10(1), pp.21–41.
- Wailoo, D.A., Tsuchiya, A. & McCabe, C., 2009. Weighting Must Wait. *PharmacoEconomics*, 27(12), pp.983–989.
- Walzer, M., 1983. *Spheres of justice: a defense of pluralism and equality*, New York: Basic Books.

- Wedel, M. et al., 1999. Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling. *Marketing Letters*, 10(3), pp.219–232.
- Weinstein, M.C., Torrance, G. & McGuire, A., 2009. QALYs: The Basics. *Value in Health*, 12, pp.S5–S9.
- Weiss, J.A., 1982. Coping with Complexity: An Experimental Study of Public Policy Decision-Making. *Journal of Policy Analysis and Management*, 2(1), pp.66–87.
- Williams, A., 1997. Intergenerational equity: an exploration of the “fair innings” argument. *Health Economics*, 6(2), pp.117–132.
- Williams, A., 1988. Priority setting in public and private health care : A guide through the ideological jungle. *Journal of health economics*, 7(2), pp.173–183.
- Williams, A., 1996. QALYs and ethics: A health economist’s perspective. *Social Science & Medicine*, 43(12), pp.1795–1804.
- Williams, A. & Cookson, R., 2000. Equity in Health. In A. J. Culyer & J. P. Newhouse, eds. *Handbook of Health Economics*. North-Holland: Elsevier, pp. 1863–1910.
- Wilmot, S. & Ratcliffe, J., 2002. Principles of distributive justice used by members of the general public in the allocation of donor liver grafts for transplantation: a qualitative study. *Health Expectations*, 5(3), pp.199–209.
- Wirtz, V., Cribb, A. & Barber, N., 2003. Understanding the role of “the hidden curriculum” in resource allocation—the case of the UK NHS. *Health Care Analysis*, 11(4), pp.295–300.
- Wiseman, V., 1997. Caring: the neglected health outcome? or input? *Health Policy*, 39(1), pp.43–53.
- Wiseman, V. et al., 2003. Involving the general public in priority setting: experiences from Australia. *Social science & medicine*, 56(5), pp.1001–1012.
- Wright, P., 1975. Consumer Choice Strategies: Simplifying Vs. Optimizing. *Journal of Marketing Research*, 12(1), pp.60–67.
- Wright, S.P., 1992. Adjusted P-Values for Simultaneous Inference. *Biometrics*, 48(4), pp.1005–1013.
- Zeileis, A. & Hothorn, T., 2002. Diagnostic Checking in Regression Relationships. *R News*, 2(3), pp.7–10.
- Zerbe, R.O. & Dively, D.D., 1994. *Benefit-cost analysis in theory and practice* 1st ed., New York: HarperCollins College Publishers.