Using discrete choice experiments to value benefits and risks in primary care

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Medical and Human Sciences

2015

Caroline Mary Vass

School of Medicine

Volume I of II
Volume I
Contents

List of Tables .............................................................................................................................................. 8
List of Figures ................................................................................................................................................ 10
List of Appendices ......................................................................................................................................... 11
Glossary .......................................................................................................................................................... 14
List of Abbreviations ..................................................................................................................................... 23
Abstract ......................................................................................................................................................... 26
Declaration ..................................................................................................................................................... 27
Copyright statement ....................................................................................................................................... 28
Acknowledgements ......................................................................................................................................... 29
The Author ...................................................................................................................................................... 30

Chapter One
Introduction ............................................................................................................................................... 31
  1.1 Thesis outline .......................................................................................................................................... 31
  1.2 Key concepts ............................................................................................................................................ 31
      1.2.1 Risk in healthcare .......................................................................................................................... 31
      1.2.2 Valuing health and healthcare ....................................................................................................... 32
  1.3 Research questions .................................................................................................................................. 32
  1.4 Evolution of research ............................................................................................................................ 32
  1.5 Study Design .......................................................................................................................................... 33
  1.6 Methods .................................................................................................................................................. 33
      1.6.1 Quantitative choice models ............................................................................................................ 33
      1.6.2 Qualitative research methods .......................................................................................................... 34
      1.6.3 Eye-tracking methods ..................................................................................................................... 34
  1.7 The case study: breast screening .......................................................................................................... 34
  1.8 Thesis outline .......................................................................................................................................... 35

Chapter Two
Why and how to elicit preferences for benefit-risk trade-offs in healthcare? ............................. 38
2.1 Introduction .................................................................................................................. 38
2.2 The role of health economics ....................................................................................... 38
   2.2.1 The role of public and patient preferences ............................................................. 39
2.3 Valuation methods ........................................................................................................ 40
   2.3.1 Revealed preferences ............................................................................................ 40
   2.3.2 Stated preference methods ..................................................................................... 42
2.4 Theoretical foundation of DCEs ................................................................................... 49
   2.4.1 Lancaster’s theory .................................................................................................. 49
   2.4.2 Random utility theory ............................................................................................ 50
2.5 Risk in healthcare valuation ......................................................................................... 52
   2.5.1 Definition of risk ................................................................................................... 52
   2.5.2 Risk in healthcare .................................................................................................. 53
   2.5.3 Risk in a DCE ....................................................................................................... 53
   2.5.4 Decisions with risk ............................................................................................... 55
      2.5.4.1 Expected utility theory .................................................................................... 56
      2.5.4.2 Risk literacy .................................................................................................... 57
   2.5.5 Communicating risk information ......................................................................... 57
2.6 Conclusion .................................................................................................................... 59

Chapter Three
Valuing risks and benefits: a case study of population-based screening for breast cancer ................................................................................................................................. 60
3.1 Introduction ................................................................................................................... 60
3.2 Risk communication formats ....................................................................................... 60
3.3 Selecting a case study ................................................................................................... 61
   3.3.1 Case study selection criteria .................................................................................. 62
   3.3.2 Identification and appraisal of case studies ............................................................. 63
3.4 Screening for breast cancer ......................................................................................... 67
   3.4.1 Background to breast screening .......................................................................... 67
   3.4.2 Facts, figures and debates ..................................................................................... 67
   3.4.3 Benefit-risks trade-offs in breast screening ............................................................ 71
3.4.4 Healthcare setting .................................................................................................................. 71
3.4.5 Potential outputs of the DCE .............................................................................................. 71
3.4.6 Existing DCEs relevant to cancer screening: a rapid review ........................................... 72
3.5 Breast screening: designing the DCE ..................................................................................... 77
  3.5.1 Research questions ........................................................................................................... 77
  3.5.2 Sampling frame ................................................................................................................ 77
  3.5.3 Identification of attributes and levels ............................................................................. 78
  3.5.4 Choice format .................................................................................................................. 83
  3.5.5 Experimental design ....................................................................................................... 83
  3.5.6 Survey materials ............................................................................................................. 86
  3.5.7 Ethical approval .............................................................................................................. 87
  3.5.8 The piloting process ........................................................................................................ 87
  3.5.9 DCE design: main study ................................................................................................. 90
3.6 Conclusion ............................................................................................................................. 93

Chapter Four

Benefit-risk trade-offs for breast screening: results of the DCE ............................................ 94

  4.1 Introduction ......................................................................................................................... 94
  4.2 Aims and objectives .......................................................................................................... 94
  4.3 Methods ............................................................................................................................. 95
    4.3.1 Data collection ............................................................................................................. 95
    4.3.2 Models ....................................................................................................................... 97
    4.3.3 Marginal rates of substitution ................................................................................... 105
  4.4 Results ............................................................................................................................... 106
    4.4.1 Descriptive statistics ............................................................................................... 106
    4.4.2 Internal validity ....................................................................................................... 107
    4.4.3 Investigating the functional form of utility ............................................................... 108
    4.4.4 Testing risk format effects ....................................................................................... 111
    4.4.5 Investigating preference heterogeneity ................................................................. 111
    4.4.6 Accounting for scale heterogeneity ....................................................................... 112
    4.4.7 Marginal rates of substitution ............................................................................... 116
Chapter Six
Benefit-risk trade-offs for breast screening: an eye-tracking study

6.1 Introduction
6.2 Research questions
6.3 Introduction to eye-tracking as a method
   6.3.1 Eye physiology and key concepts
   6.3.2 Technology and equipment
   6.3.3 Psychological foundations
   6.3.4 Research to date
6.4 Hypotheses
6.5 Methods
   6.5.1 Recruitment
   6.5.2 Apparatus
   6.5.3 Data collection
   6.5.4 Analysis
6.6 Results
   6.6.1 Participants
   6.6.2 Experimental validity
   6.6.3 Visual attention: fixation and dwell times
   6.6.4 Attribute non-attendance
   6.6.5 Saccades
   6.6.6 Pupil dilation
6.7 Discussion
   6.7.1 Key findings
   6.7.2 Strengths and limitations
6.8 Conclusion

Chapter Seven
Discussion

7.1 Introduction
7.2 Key results ...........................................................................................................210
7.3 Limitations .........................................................................................................213
7.4 Comparisons with existing literature ...............................................................217
7.5 Key implications ................................................................................................217
7.6 Future research .................................................................................................219
7.7 Conclusion .........................................................................................................220
List of Tables

Table 1.1: Chapter titles, topics covered and methods employed ........................................... 35
Table 2.1: Definitions of methods of economic evaluation .................................................. 39
Table 2.2 Components of risk ............................................................................................. 53
Table 2.3: Examples of risk in types of healthcare ................................................................. 53
Table 3.1: Potential case studies for a benefit-risk DCE ...................................................... 64
Table 3.2: Risks to women of breast cancer ........................................................................ 68
Table 3.3: Breast cancer and screening DCEs ..................................................................... 74
Table 3.4: Attributes and levels used in pilot DCE .............................................................. 82
Table 3.5: Attributes and levels used in final DCE .............................................................. 91
Table 4.1: Attributes, levels and coding .............................................................................. 97
Table 4.2: Dummy variable definitions ............................................................................... 97
Table 4.3: WTP for hypothetical screening programmes ..................................................... 106
Table 4.4: Sample characteristics ....................................................................................... 107
Table 4.5: Validity test failure and self-reported task difficulty ........................................ 108
Table 4.6: Results of the conditional logit Models 1A, 1B and 1C ...................................... 108
Table 4.7: Results of the conditional logit Model 1D .......................................................... 109
Table 4.8: Heteroscedastic conditional logit results ............................................................ 111
Table 4.9: Information criteria for varying preference-class numbers in Model 3 .......... 112
Table 4.10: Information criteria for varying class numbers in latent-class and scale-
adjusted latent-class models ............................................................................................... 113
Table 4.11: Preference-classes in scale-class one of Model 4 ............................................. 114
Table 4.12: MWTP for breast screening attributes ............................................................. 117
Table 4.13: WTP for a lifetime breast screening programme ............................................ 118
Table 4.14: WTA risk to detect an additional cancer ......................................................... 118
Table 4.15: Probability of participating in a screening programme over ‘No Screening’ .... 119
Table 5.1: Schools of thought in research methodology ....................................................... 134
Table 5.2: Interviewee characteristics ................................................................................. 147
Table 6.1: Outcomes of interest in the eye-tracking data..............................196

Table 6.2: Mean response time in milliseconds from starting the choice set to pressing a button, and differences between risk formats .................................................................197

Table 6.3: Mean number of fixations to white space per choice set, and differences between risk formats ...........................................................................................................198

Table 6.4: Visual attention to each attribute in a trial (hypothesis one) .....................201

Table 6.5: The effect of self-reported ANA on total dwell-time per trial in milliseconds (hypothesis two) .............................................................................................................202

Table 6.6: Number of choice-sets with no attention to an attribute for each risk format ..202

Table 6.7: Number and direction of saccades in a trial (hypothesis three), and difference between risk formats ...........................................................................................................203

Table 6.8: Measurements (in pixel units) of cognitive pupillometry (hypothesis four) ....204
List of Figures

Figure 2.1: Preference valuation methods ................................................................. 43
Figure 3.1: Progression of women through screening ............................................. 70
Figure 3.2: Example choice questions ..................................................................... 92
Figure 4.1: Functional form of the risk attribute for DCE respondents .................. 110
Figure 4.2: Functional form of the detection attribute for DCE respondents ........... 110
Figure 4.3: Functional form of the cost attribute for DCE respondents .................. 110
Figure 4.4: Relative importance of attributes by preference-class ....................... 115
Figure 4.5: The effect of increasing risk on uptake for breast screening by preference-class .................................................................................................................. 120
Figure 4.6: The effect of increasing detection rates on uptake for breast screening by preference-class .................................................................................................................. 120
Figure 4.7: The effect of increasing cost on uptake for breast screening by preference-class .......................................................................................................................... 121
Figure 5.1: Conversational qualitative research methods ........................................ 137
Figure 5.2: Qualitative data collection processes ..................................................... 145
Figure 5.3: Intermediate coding tree demonstrating identified codes and themes ...... 149
Figure 5.4: Final coding tree identifying key themes and inter-theme relationships .... 150
Figure 6.1: Measurement of visual degrees ............................................................... 182
Figure 6.2: Configuration of the eye-tracker .............................................................. 188
Figure 6.3: Pre-defined AOI for each choice set ....................................................... 190
Figure 6.4: Definition of direction from saccade angles ......................................... 191
Figure 6.5: Example of eye-tracking fixation data ................................................. 193
Figure 6.6: Example of collapsed fixation data to key AOI .................................... 194
Figure 6.7: Example scan path from a respondent completing the DCE with icon arrays and percentages ............................................................................................................. 199
Figure 6.8: Example scan path from a respondent completing the DCE with percentages only .......................................................................................................................... 199
Volume II

List of Appendices

Appendix 3.1: How to frame risk in a DCE: an exploration and structured review ........246
Appendix 3.2: List of studies included in the structured review of risk communication formats..............................................................................................................................................268
Appendix 3.3: The published paper containing the results of the structured review .......274
Appendix 3.4: List of DCEs with a focus on breast cancer and screening from the systematic review ........................................................................................................................................295
Appendix 3.5: The design of the DCE used in the pilot study .....................................297
Appendix 3.6: Review of questions used to elicit individual characteristics in commonly-used surveys ........................................................................................................................................300
Appendix 3.7: Confirmation letter of University Research Ethics approval .................302
Appendix 3.8: The DCE survey used for the pilot study .............................................303
Appendix 3.9: Online advertisement for study participants .......................................339
Appendix 3.10: Preliminary interview schedule for pilot study .................................340
Appendix 3.11: Information sheet for interviewees for the qualitative interview study ....341
Appendix 3.12: Consent form for interviewees for the qualitative interview study .......343
Appendix 3.13: NHS Choices video approval ...............................................................344
Appendix 3.14: Pilot quantitative results of the DCE from the internet panel .............345
Appendix 3.15: The design of the DCE used in the final empirical studies .................347
Appendix 3.16: The DCE survey used for the final empirical studies .........................349
Appendix 3.17: Quantitative results of the DCE from the internet panel in the second pilot study ........................................................................................................................................387
Appendix 4.1: ResearchNow® response rate in the large internet panel DCE .............388
Appendix 4.2: Sample characteristics of respondents who completed the DCE in the final study ........................................................................................................................................389
Appendix 4.3: Results of a simple logit regression to determine if any covariates were a significant predictor of risk communication format .................................................................392
Appendix 4.4: Latent class analysis results from 6-preference-class (no scale-classes) model for the large internet panel DCE .................................................................394
Appendix 5.1: Systematic review of qualitative research alongside healthcare DCEs .....395
Appendix 5.2: Survey to authors of published healthcare DCEs.................................420
Appendix 5.3: Flow of papers through the systematic review......................................428
Appendix 5.4: Comparison of existing appraisal tools for assessing the quality of qualitative research .................................................................................................430
Appendix 5.5: Final critical appraisal tool for assessing the quality of qualitative research alongside DCEs .........................................................................................435
Appendix 5.6: Studies included in the systematic review of qualitative research conducted alongside DCEs .........................................................................................436
Appendix 5.7: Tabulated results from the systematic review of DCEs .........................453
Appendix 5.8: List of studies not included in the Clark et al. (2014) systematic review ....455
Appendix 5.9: The submitted paper describing the development of the appraisal tool ....456
Appendix 5.10: The questionnaire sent to authors of DCE studies to assess their experience and views of using qualitative research methods ..................................................471
Appendix 5.11: The invitation to authors of DCE studies to complete the questionnaire to assess their experience and views of qualitative research methods alongside DCEs ....478
Appendix 5.12: The reminder invitation to authors of DCE studies to complete the questionnaire to assess their experience and views of qualitative research methods alongside DCEs .................................................................479
Appendix 5.13: Author responses to the questionnaire of their experience and views of conducting qualitative research alongside DCEs ..................................................480
Appendix 5.14: Paper advertisement to the general public for participation in the think-aloud interviews ............................................................................................................481
Appendix 5.15: Email advertisements to The University of Manchester staff for participation in the think-aloud interviews .................................................................482
Appendix 5.16: Think-aloud interview schedule ................................................................. 483

Appendix 5.17: Flow of recruited study participants for the think-aloud interviews and the
eye-tracking study ............................................................................................................. 485

Appendix 6.1: Paper advertisement to the general public for participation in the eye-
tracking study ..................................................................................................................... 486

Appendix 6.2: Email advertisement to The University of Manchester staff for participation
in the eye-tracking study ..................................................................................................... 487

Appendix 6.3: Information sheet for the eye-tracking study participants ......................... 488

Appendix 6.4: A description of the eye-tracking pilot study ................................................. 490

Appendix 6.5: Consent form for participants in the eye-tracking study .............................. 491

Appendix 6.6: iPad survey for the eye-tracking study ......................................................... 492

Appendix 6.7: Heteroscedastic conditional logit results for the eye-tracking study ........ 511
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative specific constant</td>
<td>A regression term representing the mean of the distribution of the unobserved effects in the random component of utility.</td>
</tr>
<tr>
<td>Anchoring effect</td>
<td>A phenomenon seen in surveys where respondents base their answers on previous responses or on the initial information provided.</td>
</tr>
<tr>
<td>Area of interest</td>
<td>Defined segments of a stimulus which can be sub-analysed by the researcher when using eye-tracking.</td>
</tr>
<tr>
<td>Attribute non-attendance</td>
<td>When respondents ignore the information contained in one attribute, violating the assumption of continuity of preferences which is key to the validation of discrete choice experiment data.</td>
</tr>
<tr>
<td>Attributes</td>
<td>The properties or characteristics of the goods from which utility may be derived and included in a discrete choice experiment.</td>
</tr>
<tr>
<td>Averting behaviour</td>
<td>Actions taken to reduce risk exposure.</td>
</tr>
<tr>
<td>Best-worst scaling</td>
<td>A type of choice experiment where respondents choose the option (alternative, attributes or levels) which yield the highest (best) and the lowest (worst) utility, revealing the components which are furthest apart on their individual utility spectrum.</td>
</tr>
<tr>
<td>Cognitive pupillometry</td>
<td>A hypothesised relationship between the size of the pupil of the eye of a respondent and their cognitive activity.</td>
</tr>
<tr>
<td>Choice set</td>
<td>The alternatives presented to the respondent from which they choose in a discrete choice experiment.</td>
</tr>
<tr>
<td>Coding tree</td>
<td>An arrangement of codes demonstrating links and potential relationships through branches from qualitative data.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Conditional logit model</td>
<td>A regression model with a categorical dependent variable where the values of the variables (usually choice characteristics) vary across the choices but parameters are common across the choices.</td>
</tr>
<tr>
<td>Conjoint analysis</td>
<td>A type of stated preference method where respondents are asked to order or score alternatives according to their preferences.</td>
</tr>
<tr>
<td>Conjoint analysis - adaptive</td>
<td>A type of stated preference method when the design is continuously updated so the choices presented to respondents depend on their previous responses.</td>
</tr>
<tr>
<td>Contingent valuation</td>
<td>A stated preference method used to elicit willingness-to-pay (or accept) values through a direct question.</td>
</tr>
<tr>
<td>Corneal reflection</td>
<td>A throwback of light from the eye which can be compared to the centre of the pupil to give an approximation of gaze.</td>
</tr>
<tr>
<td>Cost-benefit analysis</td>
<td>An assessment of the costs and the benefits of a health intervention where the benefits are measured in monetary terms.</td>
</tr>
<tr>
<td>Cost-effectiveness analysis</td>
<td>An assessment of the costs and the benefits of a health intervention where the benefits are measured in a clinical or health-related metric.</td>
</tr>
<tr>
<td>Cost-utility analysis</td>
<td>An assessment of the costs and the benefits of a health intervention where the benefits are measured in a quality-adjusted outcome, such as quality-adjusted life years.</td>
</tr>
<tr>
<td>Data reduction</td>
<td>The process of drawing meaning from large sets of qualitative data.</td>
</tr>
<tr>
<td>Data saturation</td>
<td>A principle point reached in the collection of data (often qualitative) where additional data does not contribute anything new.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>D-efficiency</td>
<td>A type of discrete choice experiment design criteria with minimal D-error satisfying, or attempting to satisfy, orthogonality, level balance, minimal overlap, and utility balance.</td>
</tr>
<tr>
<td>Design efficiency</td>
<td>Design efficiency refers to a combination of attributes within the choice sets in a discrete choice experiment, so that the maximum amount of information is obtained from respondents.</td>
</tr>
<tr>
<td>Drift</td>
<td>In eye-tracking, a gradual misalignment from a correct calibration from adjustment in the head position or a miscalculation by the tracker.</td>
</tr>
<tr>
<td>Drift correction</td>
<td>A validation procedure with an eye-tracking experiment to confirm calibration.</td>
</tr>
<tr>
<td>Dummy coding</td>
<td>A variable that takes the value of 0 or 1 to indicate the absence or presence of a category compared to a base case (the dropped variable) in a regression model.</td>
</tr>
<tr>
<td>Dwell time</td>
<td>The total time of all fixations (see definition below) of an eye to an area in an eye-tracking study.</td>
</tr>
<tr>
<td>Effects coding</td>
<td>Alternative approach to dummy variables that allow for estimation of the effects of the dropped variable.</td>
</tr>
<tr>
<td>Epistemology</td>
<td>A branch of philosophy concerned with knowledge and its limitations.</td>
</tr>
<tr>
<td>Ethnography</td>
<td>An approach, method or analysis in qualitative research that involves observing the subject.</td>
</tr>
<tr>
<td>Experimental design</td>
<td>A sample from all possible combinations of attribute levels used to construct choice alternatives in a discrete choice experiment.</td>
</tr>
<tr>
<td>Extra-welfarist</td>
<td>An evaluative framework where something other than, or in addition to, utility is maximised (for example, health).</td>
</tr>
<tr>
<td><strong>EyeLink®</strong></td>
<td>A brand of eye-tracking device used by research institutions due to their high recording speeds.</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Familiar good effect</strong></td>
<td>When the valuation of a good or service is influenced positively because it is familiar to the respondents.</td>
</tr>
<tr>
<td><strong>Fixation</strong></td>
<td>A period when the eye is ‘relatively’ still and information is processed. Typically fixations in eye-tracking are thought to last over 75 milliseconds.</td>
</tr>
<tr>
<td><strong>Fractional factorial design</strong></td>
<td>A sample from a full factorial design which can estimate effects of interest through interactions in a discrete choice experiment.</td>
</tr>
<tr>
<td><strong>Framework analysis</strong></td>
<td>An approach, method or analytical technique that uses a pre-defined set of topics to identify codes or themes from qualitative data.</td>
</tr>
<tr>
<td><strong>Full factorial design</strong></td>
<td>A design using the complete set of all attribute and level combinations in the discrete choice experiment.</td>
</tr>
<tr>
<td><strong>Gaze</strong></td>
<td>The location of focus of an eye.</td>
</tr>
<tr>
<td><strong>Generic set</strong></td>
<td>Where the labels of alternatives convey no information beyond that provided by the attributes and levels.</td>
</tr>
<tr>
<td><strong>Heteroscedastic conditional logit model</strong></td>
<td>A conditional logit regression model accounting for differences in the variance of the error term through estimation of the scale parameter.</td>
</tr>
<tr>
<td><strong>Icon array</strong></td>
<td>Also known as a risk grid, icon array is a risk communication format using individual graphics to reflect proportions most commonly out of 10, 100 or 1000.</td>
</tr>
<tr>
<td><strong>Independence of Irrelevant Alternatives</strong></td>
<td>When adding another alternative or changing the characteristics of a third alternative does not affect the relative odds between the two original alternatives in a discrete choice experiment.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Information asymmetry</td>
<td>A situation where a consumer has incomplete information because they do not possess the knowledge or expertise to assess the utility associated from a good or service.</td>
</tr>
<tr>
<td>Interpretism</td>
<td>A subjective understanding of phenomena which are deviant or unique.</td>
</tr>
<tr>
<td>Lancaster's theory</td>
<td>A hypothesis which suggests that individuals do not value a good or service <em>per se</em>, but instead value the characteristics or attributes of which it is made.</td>
</tr>
<tr>
<td>LatentGold®</td>
<td>A latent-class and finite mixture software programme used to estimate complex regression models.</td>
</tr>
<tr>
<td>Latent-class analysis</td>
<td>A regression modelling technique which identifies subsets of respondents with similar preferences.</td>
</tr>
<tr>
<td>Level balance</td>
<td>When the design of a discrete choice experiment results in the level of each attribute occurring with equal frequency in survey choice sets.</td>
</tr>
<tr>
<td>Lexicographic preferences</td>
<td>When the good providing the most of X is always preferred, no matter what the amount of Y.</td>
</tr>
<tr>
<td>Logit model</td>
<td>A probability regression model which take a logistic distribution as opposed to a normal distribution (which would be a probit).</td>
</tr>
<tr>
<td>Mammography</td>
<td>The use of X-rays of the breast(s) to create an image that can be used to diagnose cancer.</td>
</tr>
<tr>
<td>Marginal rates of substitution</td>
<td>The willingness to exchange a unit of one good for another to maintain the same level of utility.</td>
</tr>
<tr>
<td>Mixed logit model</td>
<td>A regression model which allows for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors utilising any distribution for the random coefficients.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Monotonicity</td>
<td>If A is preferred to B then A should be preferred to C if C is as good an alternative as B in all attributes.</td>
</tr>
<tr>
<td>Multinomial logit model</td>
<td>A conditional logit regression model where the values of the variables are common across choices for the same person (usually individual characteristics) but the parameters vary across choices.</td>
</tr>
<tr>
<td>Ngene®</td>
<td>A software package used to efficiently design and test the statistical properties of discrete choice experiments.</td>
</tr>
<tr>
<td>Non-use values</td>
<td>Utility derived from a good or service that is not related to current consumption.</td>
</tr>
<tr>
<td>Nonveridicality</td>
<td>When a study task or experiment is not a true reflection of real life behaviour.</td>
</tr>
<tr>
<td>NVivo®</td>
<td>A software package used in the collection, generation and analysis of qualitative data.</td>
</tr>
<tr>
<td>Ontological</td>
<td>A philosophy of study of the nature of being.</td>
</tr>
<tr>
<td>Open-coding</td>
<td>A process of identifying patterns in the qualitative data based on no prior ideas or assumptions.</td>
</tr>
<tr>
<td>Opportunity cost</td>
<td>The cost of an alternative that is forgone in making a particular choice within a fixed budget.</td>
</tr>
<tr>
<td>Optimism bias</td>
<td>A cognitive belief where an individual perceives themselves to be at less risk.</td>
</tr>
<tr>
<td>Ordering effect</td>
<td>A phenomenon which occurs when changing the arrangement of questions in a survey affects responses.</td>
</tr>
<tr>
<td>Orthogonality</td>
<td>A property of discrete choice experiments to remove collinearity and estimate level effects independently.</td>
</tr>
<tr>
<td>Payment vehicle effect</td>
<td>A phenomenon which occurs when describing a cost differently affects the values given by respondents – for example, the discrepancies between a ‘tax’ or a ‘donation’.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Positivism</td>
<td>An objective assessment of measurable absolutes.</td>
</tr>
<tr>
<td>Primacy effect</td>
<td>A response order effect where the consequence of placing a single question or choice set first. This may be because respondents use their answer to this as an anchor for responses to later sets in a discrete choice experiment.</td>
</tr>
<tr>
<td>Profile</td>
<td>A combination of attributes at varying levels which make up an option in a choice set of a discrete choice experiment.</td>
</tr>
<tr>
<td>Protest bids</td>
<td>Responses given in stated preference methods to express an opinion or to change the results of the study.</td>
</tr>
<tr>
<td>Pupillometry</td>
<td>Measure of the size of the pupil of the eye.</td>
</tr>
<tr>
<td>Qualitative research methods</td>
<td>Exploratory research to understand motivations and generate ideas or hypotheses.</td>
</tr>
<tr>
<td>Random utility theory</td>
<td>A choice theory where decisions are deterministic and utility has a random component.</td>
</tr>
<tr>
<td>Reactivity</td>
<td>Challenges for respondents to verbalise their thoughts in a cognitive interview setting.</td>
</tr>
<tr>
<td>Reflexivity (qualitative research)</td>
<td>Acknowledging the effect of a researcher on qualitative data collection.</td>
</tr>
<tr>
<td>Researcher bias</td>
<td>When social norms or the presence of someone means true thoughts/feelings are not revealed.</td>
</tr>
<tr>
<td>ResearchNow®</td>
<td>An internet panel provider who recruits samples of respondents to complete online surveys.</td>
</tr>
<tr>
<td>Revealed preference</td>
<td>Data collected through observations of behaviour in real markets.</td>
</tr>
<tr>
<td>Risk</td>
<td>The probability of a hazard (usually a negative event) occurring.</td>
</tr>
<tr>
<td>Romanticism</td>
<td>Influence of society and culture on study results.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Saccades</td>
<td>Very rapid eye movements shifting gaze to acquire information.</td>
</tr>
<tr>
<td>Sawtooth®</td>
<td>A software package, used for hosting online surveys, which can also be used to design and analyse choice experiments.</td>
</tr>
<tr>
<td>Scale-adjusted latent-class analysis</td>
<td>An extension of the latent-class regression model which identifies subsets of respondents with similar ‘randomness’ within each preference class.</td>
</tr>
<tr>
<td>Scale parameter</td>
<td>An inverse of the error variance which is a measure of systematic randomness incorporated in choice data generated from different sources. It can also be thought of as a measure of choice consistency.</td>
</tr>
<tr>
<td>Stability (preferences)</td>
<td>If A is preferred to B at one point, A should still be preferred to B later.</td>
</tr>
<tr>
<td>Starting-point bias</td>
<td>A phenomenon in contingent valuation where respondents are influenced by the first question or bid.</td>
</tr>
<tr>
<td>STATA®</td>
<td>A software package used for the statistical analysis of quantitative data.</td>
</tr>
<tr>
<td>Stated preference</td>
<td>Data collected through surveying individuals to attain information about how they would behave in the hypothetical scenario presented.</td>
</tr>
<tr>
<td>Stimulus</td>
<td>In eye-tracking, this is the object that the participant is likely to react to or see.</td>
</tr>
<tr>
<td>Structuring</td>
<td>Grouping themes in qualitative analysis to make sense of the data.</td>
</tr>
<tr>
<td>Trial</td>
<td>A test in an experiment. In the context of eye-tracking, each trial refers to each recorded screen an individual sees.</td>
</tr>
<tr>
<td>Thematic analysis</td>
<td>The generation of clusters of codes from qualitative data related to similar topics.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>Think-aloud (verbal protocol analysis)</td>
<td>Technique used in qualitative research and cognitive interviewing to attain information on individuals' thought processes. Individuals are asked to verbalise their thoughts as they complete a task.</td>
</tr>
<tr>
<td>Trade-off</td>
<td>A compromise which minimises the opportunity cost to maximise benefit or utility.</td>
</tr>
<tr>
<td>Transferability</td>
<td>A qualitative measure of generalisability.</td>
</tr>
<tr>
<td>Transitivity (preferences)</td>
<td>If A is preferred to B, and B to C, then A is preferred to C.</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>Where the probability of an event occurring is unknown because it is ambiguous or because there are surrounding confidence intervals.</td>
</tr>
<tr>
<td>Utility</td>
<td>Term used in economics to describe the satisfaction gained from the consumption of goods or services.</td>
</tr>
<tr>
<td>Utility balance</td>
<td>A design criterion satisfied when the estimated utility of each alternative within a choice set in a discrete choice experiment is the same.</td>
</tr>
<tr>
<td>Visual attention</td>
<td>Eye-movements which result in fixations that indicate an object is being considered by an individual.</td>
</tr>
<tr>
<td>Visual degrees</td>
<td>A measure of distance in eye-tracking which allows for varying proximity between the object and pupil.</td>
</tr>
<tr>
<td>Warm-glow theory</td>
<td>Utility acquired from altruistic or charitable behaviour.</td>
</tr>
<tr>
<td>Welfarist</td>
<td>An evaluative framework that maximises total utility for society.</td>
</tr>
<tr>
<td>Yea-say bias</td>
<td>A tendency to agree with what was being asked (either in a survey or interview).</td>
</tr>
<tr>
<td>Zero bids</td>
<td>A phenomenon present in contingent valuation where a bid of nil is given for something the respondent does actually value.</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Term</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>ACA</td>
<td>Adaptive Conjoint Analysis</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>AMD</td>
<td>Age-Related Macular Degeneration</td>
</tr>
<tr>
<td>ANA</td>
<td>Attribute Non-Attendance</td>
</tr>
<tr>
<td>AOI</td>
<td>Areas of Interest</td>
</tr>
<tr>
<td>ASC</td>
<td>Alternative Specific Constant</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>BMA</td>
<td>British Medical Association</td>
</tr>
<tr>
<td>BMI</td>
<td>Body Mass Index</td>
</tr>
<tr>
<td>BWS</td>
<td>Best-Worst Scaling</td>
</tr>
<tr>
<td>CAIC</td>
<td>Consistent Akaike Information Criterion</td>
</tr>
<tr>
<td>CBA</td>
<td>Cost-Benefit Analysis</td>
</tr>
<tr>
<td>CEA</td>
<td>Cost-Effectiveness Analysis</td>
</tr>
<tr>
<td>CJA</td>
<td>Conjoint Analysis</td>
</tr>
<tr>
<td>CRD</td>
<td>Centre for Reviews and Dissemination</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-Separated Value</td>
</tr>
<tr>
<td>CUA</td>
<td>Cost-Utility Analysis</td>
</tr>
<tr>
<td>DC</td>
<td>Dummy Coding</td>
</tr>
<tr>
<td>DCE</td>
<td>Discrete Choice Experiment</td>
</tr>
<tr>
<td>DT</td>
<td>Dwell Time</td>
</tr>
<tr>
<td>EC</td>
<td>Effects Coding</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>EuroQol Five Dimension</td>
</tr>
<tr>
<td>EUT</td>
<td>Expected Utility Theory</td>
</tr>
<tr>
<td>GP</td>
<td>General Practitioner</td>
</tr>
<tr>
<td>Hz</td>
<td>Hertz</td>
</tr>
<tr>
<td>IAP</td>
<td>Icon Arrays and Percentages</td>
</tr>
<tr>
<td>ID</td>
<td>Identity</td>
</tr>
<tr>
<td>IIA</td>
<td>Independence of Irrelevant Alternatives</td>
</tr>
<tr>
<td>ISPOR</td>
<td>International Society for Pharmacoeconomics and Outcomes Research</td>
</tr>
<tr>
<td>LL</td>
<td>Log Likelihood</td>
</tr>
<tr>
<td>LR</td>
<td>Likelihood Ratio</td>
</tr>
<tr>
<td>MF</td>
<td>Mean Fixation Duration</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation</td>
</tr>
<tr>
<td>MNL</td>
<td>Multinomial Logit</td>
</tr>
<tr>
<td>MRS</td>
<td>Marginal Rate of Substitution</td>
</tr>
<tr>
<td>ms</td>
<td>Millisecond</td>
</tr>
<tr>
<td>MWTP</td>
<td>Marginal Willingness-to-Pay</td>
</tr>
<tr>
<td>MXL</td>
<td>Mixed Logit</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>NHSBSP</td>
<td>National Health Service Breast Screening Programme</td>
</tr>
<tr>
<td>NICE</td>
<td>National Institute for Health and Care Excellence</td>
</tr>
<tr>
<td>NIHR</td>
<td>National Institute for Health Research</td>
</tr>
<tr>
<td>NNT</td>
<td>Number Needed to Treat</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>NOF</td>
<td>Number of Fixations</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary Least Squares</td>
</tr>
<tr>
<td>ONS</td>
<td>Office for National Statistics</td>
</tr>
<tr>
<td>Pc</td>
<td>Preference-class</td>
</tr>
<tr>
<td>PO</td>
<td>Percentages Only</td>
</tr>
<tr>
<td>PRISMA</td>
<td>Preferred Reporting Items for Systematic Reviews and Meta-Analyses</td>
</tr>
<tr>
<td>QALY</td>
<td>Quality-Adjusted Life Year</td>
</tr>
<tr>
<td>RCT</td>
<td>Randomised Controlled Trial</td>
</tr>
<tr>
<td>RT</td>
<td>Response Time</td>
</tr>
<tr>
<td>RUT</td>
<td>Random Utility Theory</td>
</tr>
<tr>
<td>Sc</td>
<td>Scale-class</td>
</tr>
<tr>
<td>SPCR</td>
<td>School for Primary Care Research</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>VAS</td>
<td>Visual Analogue Scale</td>
</tr>
<tr>
<td>WTA</td>
<td>Willingness-to-Accept</td>
</tr>
<tr>
<td>WTP</td>
<td>Willingness-to-Pay</td>
</tr>
</tbody>
</table>
Abstract

The University of Manchester

Caroline Mary Vass

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Medical and Human Sciences

Using discrete choice experiments to value benefits and risks in primary care

2015

Discrete choice experiments (DCEs) are a stated preference valuation method. As a ubiquitous component of healthcare delivery, risk is increasingly used as an attribute in DCEs. Risk is a complex concept that is open to misinterpretation; potentially undermining the robustness of DCEs as a valuation method. This thesis employed quantitative, qualitative and eye-tracking methods to understand if and how risk communication formats affected individuals’ choices when completing a DCE and the valuations derived.

This thesis used a case study focussing on the elicitation of women’s preferences for a national breast screening programme. Breast screening was chosen because of its relevance to primary care and potential contribution to the ongoing debate about the benefits and harms of mammograms. A DCE containing three attributes (probability of detecting a cancer; risk of unnecessary follow-up; and cost of screening) was designed. Women were randomised to one of two risk communication formats: i) percentages only; or ii) icon arrays and percentages (identified from a structured review of risk communication literature in health).

Traditional quantitative analysis of the discrete choices made by 1,000 women recruited via an internet panel revealed the risk communication format made no difference in terms of either preferences or the consistency of choices. However, latent class analysis indicated that women’s preferences for breast screening were highly heterogeneous; with some women acquiring large non-health benefits from screening, regardless of the risks, and others expressing complete intolerance for unnecessary follow-ups, regardless of the benefits. The think-aloud method, identified as a potential method from a systematic review of qualitative research alongside DCEs, was used to reveal more about DCE respondents’ decision-making. Nineteen face-to-face cognitive interviews identified that respondents felt more engaged with the task when risk was presented with an additional icon array. Eye-tracking methods were used to understand respondents’ choice making behaviour and attention to attributes. The method was successfully used alongside a DCE and provided valid data. The results of the eye-tracking study found attributes were visually attended to by respondents most of the time.

For researchers seeking to use DCEs for eliciting individuals’ preferences for benefit-risk trade-offs, respondents were more receptive to risk communicated via an icon array suggesting this format is preferable. Policy-makers should acknowledge preference heterogeneity, and its drivers, in their appraisal of the benefits of breast screening programmes. Future research is required to test alternative risk communication formats and explore the robustness of eye-tracking and qualitative research methods alongside DCEs.
Declaration

This dissertation is entirely the result of my own research. No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses
Acknowledgements

I would first like to express my gratitude to my supervisory team: Katherine Payne, Dan Rigby and Stephen Campbell for their intellectual support and guidance throughout this PhD. I would like to thank Dan for his assistance with the development of my quantitative skills. I would particularly like to thank Katherine for her extensive feedback, constructive criticism, patience, time and moral support whilst completing this thesis.

I would like to thank the National Institute for Health Research School for Primary Care Research for providing the funding for this PhD. The monies gave me the opportunity to attend conferences to acquire feedback on my work and develop networks within the field of health economics. Funding for training and courses was key in developing the mixed – methods skills required to complete this thesis.

I would also like to thank my research participants for taking time out of their lives to provide data for this thesis.

I would like to thank my family: Mum, Dad, Pete, Tom, Olive and EJ for only being a phone call away, and cheering me up constantly.

I would finally like to thank Alex Turner for being there for me whenever I needed him, and keeping me fed, watered and loved.
**The Author**

The author, Caroline, was awarded a Bachelor of Science in Economics degree (Class 2 Division I (Hons)) from the University of Birmingham in July 2010. She was awarded a Master of Science in Health Economics and Health Policy (Merit) also from the University of Birmingham in December 2011.

In October 2011, she was awarded a National Institute for Health Research School for Primary Care Research doctoral trainee studentship to complete this PhD at the Manchester Centre for Health Economics, The University of Manchester. In October 2014, Caroline won the Lee B Lusted Student Prize at the 36th Annual Meeting of the Society for Medical Decision Making after presenting pilot work conducted in this PhD thesis.
Chapter One

Introduction

1.1 Thesis outline
The overall aim of this thesis was to identify and quantify benefit-risk trade-offs when eliciting preferences for healthcare interventions with discrete choice experiments (DCEs). This PhD used a combination of quantitative (DCEs), qualitative (think-aloud) and eye-tracking methods to understand individuals’ preferences for benefit-risk trade-offs. A ‘benefit-risk trade-off’ was defined as the balance between a gain (to health) and the probability of incurring harm. The primary outcome of interest was how DCE respondents’ choices and decision strategies were affected by the format of the risk attributes.

This chapter introduces the broad topic of risk in healthcare and briefly describes the selected valuation method, DCEs, before outlining the research questions in sections 1.2 and 1.3, respectively. The evolution of the research programme for this thesis is explained in section 1.4 and in section 1.5 the studies used in the thesis are outlined. The methods used in the thesis are introduced in section 1.6. In section 1.7, the case study of the United Kingdom’s (UK) National Health Service (NHS) breast screening programme (NHSBSP) is described. Section 1.8 outlines the structure of the thesis.

1.2 Key concepts
This section introduces the key terms and concepts used in this PhD. All definitions are explained in more detail in subsequent chapters. In section 1.2.1, risk and its role in healthcare is briefly introduced. In section 1.2.2, the valuation method used in the thesis, DCEs, is briefly described.

1.2.1 Risk in healthcare
Risk is the probability of an event (often negative) occurring. Most choices in life involve an element of risk; the probability of a traffic accident on the way to work, or risking food poisoning for a ‘cheap-eat’. Similarly, choices in healthcare often involve an element of risk; this could be in the form of a side-effect from an interaction or the risk of misdiagnosis. Many of the choices made in real-life are carried out subconsciously without explicit statements of the probability of particular events occurring. For this, and other reasons, it is well established that people find the concept of risk and probability difficult to understand (Fagerlin et al., 2007). It is this problem that marks the starting point for this thesis. It is known that risk is a poorly understood concept by many individuals (Lipkus et
al., 2001), but this thesis is particularly concerned with how this impacts on the valuations of healthcare goods and services derived by health economists using DCEs.

1.2.2 Valuing health and healthcare
Health economics, in general, aims to use economic theories and methods to inform the efficient allocation of finite healthcare resources (Drummond & McGuire, 2001). Therefore decision-makers need to consider which intervention or service generates the most benefit for populations for a given cost. DCEs are an increasingly popular valuation method used to measure the benefit of a healthcare good or service (de Bekker-Grob et al., 2012; Clark et al., 2014). In a DCE, individuals are presented with a survey comprising hypothetical goods or services, described by their attributes, and asked to select their preferred alternative in a series of choice sets. The valuation method estimates the marginal rates of substitution (see glossary for a definition) from the choices made, and can, therefore, be useful in eliciting people’s preferences for benefit-risk trade-offs (Johnson et al., 2010; Hauber et al., 2010).

1.3 Research questions
Four research questions are specifically related to the stated aim of this thesis (see p.31):

1. To what extent is qualitative research reported in health DCEs and how useful are these research methods?
2. What is the most appropriate risk communication format to use in public surveys?
3. How sensitive are valuations from a DCE to the methods used in communicating risk?
4. How effective are alternatives to qualitative methods in evaluating respondents’ understanding of attributes and their levels?

The overall aim and research questions were developed in response to the increased use of DCEs for eliciting preferences for benefit-risk trade-offs. The research questions were devised to robustly investigate the overall aim by utilising, and assessing the usefulness, of different research methods.

1.4 Evolution of research
The approaches and methods conducted as part of this PhD were developed concurrently with the research. From the outset it was unclear which methods would be appropriate for understanding the effects of alternative risk formats on choice behaviour in a DCE. During the course of this PhD, the research questions and research methods evolved substantially. Whilst exploratory reviews were conducted in the early phases of the PhD, a subsequent
systematic review of DCEs informed the appropriate qualitative methods and analytical perspective, in addition to aiding the selection of an original application of the DCE task (breast screening), and identifying a format for the framing of risk attributes.

1.5 Study Design
To investigate the impact of the framing of risk attributes in a DCE, the thesis was split into two components: 1) a synthesis of the relevant literatures; and 2) empirical experiments. The first component involved: 1) a literature review of valuation methods in health economics and economics more generally; 2) a systematic review of published DCEs with a particular focus on the qualitative methods employed by researchers; and 3) a structured review of risk communication formats with the aim of understanding the gold-standard of risk communication formats. These literature reviews were combined to inform the case study and DCE design. Three empirical experiments were then conducted: 1) a large sample DCE with discrete choice modelling and quantitative interpretation of the results; 2) a think-aloud study involving face-to-face interviews with DCE respondents and qualitative data analysis; and 3) an eye-tracking study where DCE respondents’ choice making behaviour was measured through their visual attention and attendance. These three empirical pieces of work were compared with one another and then bought together in a discussion of the implications of their collective findings (see Chapter Seven).

1.6 Methods
This PhD used mixed methods to answer the four research questions (described in section 1.3). The specific objectives for each component of the study are stated at the start of the relevant chapter. To achieve the stated objectives, different research methods were employed to provide a mechanism for subsequent triangulation of the results. The three broad methods used were: 1) estimation of random utility choice models to derive marginal utilities and the marginal rates of substitution between different attributes; 2) qualitative research methods (think-aloud) to acquire a self-reported account of the cognitive processes involved in the completion of a DCE; and 3) eye-tracking methods to monitor respondents’ attention and attendance to attributes in choice sets in the DCE.

1.6.1 Quantitative choice models
It is generally assumed that DCE respondents make choices which would maximise their utility. Random utility theory (RUT) is the economic theory which supports this assumption, but also allows for a random component which could be due to these measurement errors, latent attributes or unobservable taste heterogeneity. It is impossible to say with certainty which alternative a respondent will choose, therefore, probabilistic
discrete choice modelling is used to explain the uncertainty around predicting respondents’ choices. People’s preferences for healthcare have been shown to be highly heterogeneous (Hole, 2008). Latent-class models aim to estimate groups of respondents with similar decision rules, which may provide a more useful interpretation for decision-makers (Louviere 2006). The empirical quantitative chapter (see Chapter Four) of the thesis estimates models with different numbers of latent-classes.

1.6.2 Qualitative research methods
In general, health economists continue to use the quantitative methods used in mainstream economics. This is despite calls for the use of qualitative research methods in health economics (Coast, 1999; Coast et al., 2004). The potential usefulness of qualitative research methods, particularly in the area of DCEs, has been acknowledged (Kløjgaard et al., 2012; Coast & Horrocks, 2007; Coast et al., 2012) and is somewhat incorporated in guidelines for the design, conduct and interpretation of DCEs (Lancsar & Louviere, 2008; Bridges et al., 2011a). In general, the focus in the literature to date has been to use qualitative research methods to design the DCE. In contrast, this PhD seeks to use qualitative research methods to understand more about how a DCE was completed and how the formatting of risk attributes may aid, or hinder, choice making. The qualitative empirical study in this thesis (see Chapter Five) used cognitive ‘think-aloud’ interviews (a form of verbal protocol analysis). The qualitative data were then analysed using an approach derived from a form of open-coding and framework analysis.

1.6.3 Eye-tracking methods
For many years researchers in psychology have used eye-tracking to investigate cognitive processes (Rayner, 1998). The method is well established in psychology and there have also been published studies in the fields of marketing and consumer choice (Bialkova & van Trijp, 2011). The method is supported by the ‘eye-mind hypothesis’ suggested by Just & Carpenter (1980) which provides the underpinning to most psychological analyses of eye-tracking data. However, there are limited examples of eye-tracking being used alongside DCEs. In this thesis, the use of eye-tracking as a method was exploratory and aimed to understand if the technique and results could be potentially useful to interpret choice making and impact of the formatting of risk in DCEs.

1.7 The case study: breast screening
The case study used in this thesis involved valuing preferences for a national breast screening programme. Screening for breast cancer using mammograms has been proven to detect cases of breast cancer earlier (NHS Information Centre Screening and
Immunisations 2012). Women who participate in screening programmes have been shown to have improved mortality rates because of earlier subsequent intervention (Independent UK Panel on Breast Cancer Screening, 2012). However, there is a risk that a mammogram will result in a woman being recalled for biopsies and unnecessary tests because of a false-positive (Welch & Black, 2010). The unnecessary follow-ups are not only painful; an incorrect test result can cause undue worry for the woman and her family (Johnston et al., 1998).

In England, the NHSBSP invites all women to be screened between the ages of 50 to 70 years. The breast screening programme prevents an estimated 1500 deaths a year and costs an estimated £96 million annually (Pharoah et al., 2013; Public Health England, 2013a). However, the benefits of breast screening, given the risks of unnecessary follow-up and the over-diagnosis of slow-growing cancer, have been extensively debated in leading journals such as the British Medical Journal and The Lancet (Gøtzsche & Nielsen, 2009; Gøtzsche & Olsen, 2000; Baum, 2013). As a consequence of the debate, a complete review of the UK’s screening policy was undertaken in 2010 by the Independent UK Panel on Breast Cancer Screening who published their report in 2012. Quantification of women’s preferences for breast screening could make a useful contribution to the discussion of the potential balance between risks and benefits.

1.8 Thesis outline

This chapter has introduced the key concepts and methods used in the thesis. The subsequent chapters in this thesis present reviews, methods (a DCE for breast screening), and empirical studies. An outline of the topics and methods used in each chapter of the thesis are described in Table 1.1. In addition, a glossary and a list of abbreviations are presented at the start of the thesis to aid understanding of the economic, clinical and psychological terminology.

### Table 1.1: Chapter titles, topics covered and methods employed

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Key Methods</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two</td>
<td>Literature review</td>
<td>Why and how to elicit preferences for benefit-risk trade-offs in healthcare</td>
</tr>
<tr>
<td>Three</td>
<td>Efficient experimental design; Pilot interviews</td>
<td>Valuing risks and benefits: a case study of population-based screening for breast cancer</td>
</tr>
<tr>
<td>Four</td>
<td>Discrete choice modelling</td>
<td>Benefit-risk trade-offs for breast screening: results of the DCE</td>
</tr>
<tr>
<td>Five</td>
<td>Qualitative interviews</td>
<td>Benefit-risk trade-offs for breast screening: a qualitative study</td>
</tr>
<tr>
<td>Six</td>
<td>Eye-tracking trials</td>
<td>Benefit-risk trade-offs for breast screening: an eye-tracking study</td>
</tr>
<tr>
<td>Seven</td>
<td>Assimilation of findings from reviews and empirical studies</td>
<td>Discussion</td>
</tr>
</tbody>
</table>
The thesis starts with a general literature review presented in Chapter Two, which introduces the underlying theories, concepts and methods in economic valuation. Whilst focussed on health, this chapter draws upon examples and research from environmental, transport and food economics. The chapter describes why DCEs are an appropriate valuation method for eliciting preferences for benefit-risk trade-offs, but the challenges of communicating risk in a choice setting.

Chapter Three explains why, and how, breast screening was selected as a case study for this PhD. The chapter reviews existing screening DCEs identified by the systematic review. Chapter Five also describes the development and design of the DCE survey, including the identification of the two risk communication formats via a structured review. The iterative stages of piloting the survey with patients, experts, and the public are also explained. The chapter finishes by describing the final DCE used in the subsequent empirical chapters (see Chapters Four, Five and Six) of this thesis. Two versions of the DCE were created: version 1) risk attributes communicated as percentages only; and version 2) risk attributes communicated as percentages and as an icon array. These two versions were used as the basis for comparing the impact of different risk communication formats in a DCE.

Chapter Four details a study of a large sample online DCE which was administered to female members of the public recruited via an internet panel. The DCE designed in Chapter Three was completed by a sample of women randomised to one of the two risk communication formats (risk communicated with percentages only; or with percentages and an icon array). The chapter explains the econometric specification of discrete choice models and the final analytical approach taken. The quantitative results of the large DCE study are presented in terms of the women’s benefit-risk trade-offs and willingness-to-pay (WTP).

Chapter Five presents a qualitative study using the think-aloud method. The ‘think-aloud’ method was identified from a systematic review of qualitative research methods alongside DCEs and a survey to the authors of these DCEs to understand the usefulness of such methods. The DCE designed in Chapter Three was fielded in a sub-sample of women randomised to one of the two risk communication formats (risk communicated with percentages only; or with percentages and an icon array). The results of the think-aloud interviews are presented with verbatim quotes from the interviewees.

Chapter Six describes eye-tracking as a method and explains how it could be useful in understanding DCE respondents’ decision strategies. The chapter reports the methods and
results of an eye-tracking study which used the DCE designed in Chapter Three with a further sub-sample of women randomised to one of the two risk communication formats (risk communicated with percentages only; or with percentages and an icon array).

An overview of the findings of the empirical research (conducted in Chapters Four, Five and Six) are summarised and discussed in Chapter Seven. This chapter also considers the implications of the findings of this research for other DCE practitioners before suggesting future research topics and potential methods to use.

This chapter has briefly outlined and explained the rationale behind the research conducted in the thesis. The next chapter, Chapter Two, builds upon the key ideas introduced here and describes the foundations of DCEs within health economics decision-making and economic theory.
Chapter Two
Why and how to elicit preferences for benefit-risk trade-offs in healthcare?

2.1 Introduction
The aims of Chapter Two are to describe: 1) why it is important to elicit preferences for healthcare interventions; 2) the methods available to elicit preferences; and 3) the challenges of using the stated preference method, DCEs, for valuing benefit-risk trade-offs. The Chapter presents an overview of the relevant literature to address these aims. Formal systematic review methods were not used but the identified literature provides an extensive overview of the current evidence on applied research and methodological challenges when designing DCEs.

2.2 The role of health economics
Health resources are finite because decisions, such as those made by health service commissioners or policy makers, are subject to a strict budget constraint set by the available annual healthcare budget (Appleby, 2013). Within a finite budget, the funding of any healthcare intervention or service creates an opportunity cost (Gold et al., 1996). This opportunity cost arises as another area of expenditure must be reduced or sacrificed completely. Efficiency is achieved through allocating resources to produce the most benefit, but also through ensuring healthcare production is achieving an optimal output given the resources consumed (Retzlaff-Roberts et al., 2004).

In the UK, decision-making bodies such as the National Institute for Health and Care Excellence (NICE) use information from economic evaluations to inform the efficient allocation of healthcare resources (Drummond & McGuire, 2001). Table 2.1 defines the available methods of economic evaluation.
### Table 2.1: Definitions of methods of economic evaluation

<table>
<thead>
<tr>
<th>Type of economic evaluation</th>
<th>Definition (from Drummond &amp; McGuire (2001))</th>
<th>Theory underpinning the method*</th>
<th>Applied example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost-effectiveness analysis (CEA)</td>
<td>An evaluation where the costs of a health technology or service are weighed against outcomes measured in units that are readily counted, such a life years gained.</td>
<td>Extra-welfarist</td>
<td>Puri et al. (2012) compared radiotherapy to surgery in the treatment of lung cancer through the primary outcome, life-years gained.</td>
</tr>
<tr>
<td>Cost-utility analysis (CUA)</td>
<td>An evaluation where benefit is measured using a combination of health state and length of life, for example, the Quality Adjusted Life Year (QALY).</td>
<td>Extra-welfarist</td>
<td>Athanasakis et al. (2012) evaluated a hypertension intervention using QALYs to compare treatment as opposed to no intervention at all.</td>
</tr>
<tr>
<td>Cost-benefit analysis (CBA)</td>
<td>An evaluative framework that elicits people’s valuation of benefit for a good or service using ‘money’ as the unit of measurement and compares this to the associated costs.</td>
<td>Welfarist</td>
<td>Muangchana &amp; Warinsatian (2011) evaluated an influenza vaccination programme using WTP values elicited through an online survey with the specific aim of capturing the wide range of intangible benefits from the vaccination.</td>
</tr>
</tbody>
</table>

* Welfarist is defined as maximising utility to individuals; extra-welfarist is defined as maximising something other than, or in addition to, utility (for example, individuals’ health).

The aim of using evidence generated from economic evaluations is to minimise opportunity costs or maximise benefits (Gafni & Birch, 2003). Therefore decision-makers allocating finite healthcare resources need to consider which intervention or service creates the most benefit. Methods of economic evaluations compare the total benefits with the total costs of a new intervention or technology to current practice thus indicating where resources are best allocated (NICE, 2013).

NICE currently takes an extra-welfarist view and uses model-based cost-effectiveness analysis to inform decisions. Model-based cost-effectiveness analysis involves comparing the relative costs of a new intervention with existing interventions with gains in health status, quantified in terms of extension to, and the quality of life, by using the QALY (NICE, 2013). Cost-benefit analyses, underpinned by the welfarist view, have yet to be used to inform national resource allocation decisions in the UK and this is largely down to problems with the methodologies used to elicit WTP values (Grutters et al., 2008; Cookson, 2003).

#### 2.2.1 The role of public and patient preferences

There have been increased calls for more patient and public involvement in healthcare decision-making at all levels (Hogg, 2007; Mockford et al., 2011; Stewart et al., 2011). Patient involvement is important at an individual level (shared decision-making), policy level (NICE patient experts) and at a commissioning level (incorporating patient preferences in technology evaluations). As well as occurring at different levels, patient
involvement can occur to varying degrees. For example, patients may play a crucial role in developing the direction of a research project or they may provide vital data on their utility for health states (Boote et al., 2010; Brazier et al., 1999). However, in order to incorporate public and patients’ views into healthcare decision-making, the methods used must be robust, underpinned by theory, and able to accurately elicit preferences (Cookson, 2003).

2.3 Valuation methods
Methods used to value preferences for healthcare services and interventions are generally split into revealed and stated preference methods.

2.3.1 Revealed preferences
Revealed preference data are acquired through direct observation of consumer behaviour and analysis of real demand curves through observable changes in existing markets (Ryan & Gerard, 2003a). Therefore to use revealed preference data, an observable market must exist.

2.3.1.1 Travel cost method
One revealed preference method available for valuing a good or service is the travel cost method (Hanley et al., 2001). Under this approach, the costs incurred by an individual when they choose to visit an amenity or service, such as their cost of travel or loss of work, are said to be indicative of their valuation (Hanley et al., 2001). The approach has been used extensively within environmental economics where data are regularly published (Willis & Garrod, 1991; Chen et al., 2004; Font, 2000). As an example, Blakemore & Williams (2008) used the travel cost method to value a beach in Turkey taking into account British tourists’ expenditure on flights, accommodation and meals as well as the opportunity cost of spending their time there as opposed to enjoying other activities. Although the travel cost method is less popular in health economics, published health-related studies have looked at access costs to clinics to estimate patients’ WTP for more local service provision (Clarke, 2002). However, the method has been criticised because of its superficiality and reliance on an accurate valuation of the cost of time, the exclusion of equipment costs, and the difficulty in distinguishing what is valued by consumers (Randall, 1994).

2.3.1.2 Hedonic pricing
Another approach used to measure revealed preferences is the hedonic pricing method, which involves assessing the market-value of a particular good or service through price premiums or discounts (Perman et al., 2003). Studies in health economics, Jensen &
Morrisey (2011) and Robst (2006), have considered the demand for various health insurance policies and the trade-offs consumers were willing to make between premiums and cover. Market failure in healthcare means prices are rarely present and it can be difficult to find data to support this method particularly in a publically-provided healthcare system such as the UK (McIntosh et al., 2010).

2.3.1.3 Averting behaviour

Behaviour revealed in wage premiums accepted by workers for riskier jobs, or individuals’ averting behaviour through their demand for risk-reducing safety measures in vehicles such as airbags or seat-belts can also be used to measure revealed preferences (Rohlfs et al., 2015; Hakes & Viscusi, 2007; Perman et al., 2003). For example, Bresnahan et al. (1997) looked at the costs involved in either averting or mitigating dangers to value clean air in Los Angeles with measures of the steps taken to prevent smog inhalation and behavioural reactions to ‘bad-air days’ considered to indicate value. However, difficulties in distinguishing between people who do not change their behaviour in response to risks because they have permanently abstained from the risky activity and those who do not value the activity at all were acknowledged by Bresnahan et al. (1997) as creating bias in the results.

Researchers can, in general, be confident that revealed preference data are reflecting actual behaviour but the data are not always available or useful for valuations of healthcare interventions or services (Ryan et al., 2008). In healthcare, the goods or services being valued may not presently be offered and, therefore, revealed preference methods can only be used for auditing current policy, rather than informing new policy or proposed changes to the service (Ryan et al., 2008). In addition, the revealed preference data are often difficult to identify in the UK, where patients rarely face market conditions when making decisions about healthcare consumption (McIntosh et al., 2010).

There are two key challenges with using revealed preference methods in the context of healthcare: 1) decision-makers often require information on a new service or a service for which people may not have prior experience; and 2) the need to apply the principles of opportunity cost means we need to understand what, if anything, was ‘given up’, partially forgone or not chosen (Ryan et al. 2008). Another common criticism, particularly within the environmental literature, is that revealed preference data do not allow for non-use values (Cicchetti & Wilde, 1992). For example, despite no observational data on some people accessing a clinic, these individuals may still derive value from its availability for others or for their own future consumption.
Limited and absent data from revealed preferences has resulted in the development of stated preference methods which may be preferable in the context of public goods such as healthcare (Train, 2009). Section 2.3.2 describes available stated preference methods and how they overcome some of the criticisms of the revealed preference methods described in section 2.3.1.

### 2.3.2 Stated preference methods

Stated preference methods can be used to elicit preferences to generate values of monetary benefit, for use in a CBA, or values in terms of preference weights (to generate utility values) associated with a health state, for use in a CUA. Stated preference methods involve asking people to state how they would behave, or what they would prefer, based on a hypothetical scenario (Louviere et al. 2000). Humans make choices on a daily basis. Given that humans live in a choice-based world, one suggested approach to classifying the available stated preference methods is to group them into trade-off based and non-choice approaches with a strong recommendation to use trade-off based methods which reflect day-to-day human decision-making behaviour (Carson et al., 2009). Categorisation using system is not based on theory per se but is a practical approach that has been adopted by a number of authors (see: Carson et al. 2009; Louviere & Islam 2008). Figure 2.1 was developed to explain different preference methods after a comprehensive review of the stated preference literature. Figure 2.1 is used to inform sections 2.3.2.1 to 2.3.2.5 which describe the available types of stated preference methods.
Figure 2.1: Preference valuation methods

Adapted from Carson & Louviere (2011)
2.3.2.1 Non-choice methods

The visual analogue scale (VAS) is a non-choice based approach and could be used to generate a utility value for use in CUA (Drummond & McGuire, 2001). The VAS method has been used to elicit individuals’ preferences for different health states often by providing a scale between 0, death, and 1, perfect health (Whitehead & Ali, 2010). Respondents then rate various health states which can be used to estimate the weightings they place on each state and transform length of life into QALYs (Parkin & Devlin 2006). A fundamental criticism of the VAS is that it does not consider opportunity costs as the respondent is not required to trade or give-up anything when rating the health states (Brazier et al., 1999), and therefore VAS is generally not recommended as a method for eliciting utility values.

Other methods, which align with the concept of opportunity cost, include the trade-off based stated preference methods. In trade-off based methods respondents are asked to trade and make decisions about value by exchanging health for time, money, risk or other good-related attributes. Whole service trade-off based stated preference methods tend to be used to value ‘whole services’ but, in principle, there is no reason why different attributes of a service could be valued in separate surveys. In practice, this step-wise approach to valuing service attributes is not done due to potential problems with cognitive overload for survey respondents, resulting in strategic bias (such as task learning) (Holmes & Boyle, 2004).

2.3.2.2 Standard gamble

Standard gamble is used to elicit utility values for use in CUA and involves asking people to disclose how much risk, quantified using a probability, that they are prepared to accept to undergo a treatment that will either cure them from a health state or cause death (Jones, 2007). The less desirable the health state, the more risk is taken and therefore it is possible to estimate the relationship between different levels of quality of life and provide weightings for QALYs (Drummond & McGuire, 2001). It is well established that people find risk a difficult concept to understand and, although trade-offs are made in everyday life, explicitly stated probabilities and changes in magnitudes have been difficult to interpret for standard gamble respondents (Dolan & Iadarola 2008). The concept relies heavily on the assumption that these differences in magnitude are appreciated by the respondent and therefore the method is highly susceptible to framing effects (Abellan-Perpinan et al., 2005).
2.3.2.3 Time trade-off

The time trade-off is a method also used to elicit utility values for use in CUA (Drummond & McGuire, 2001). The time trade-off method was developed specifically to value health states and it is not based on the economic utility theory that supports the standard gamble method (Bleichrodt & Johannesson, 1997). In the time trade-off method, respondents are required to trade length of life for quality of life (Torrance, 1986). For example, respondents may be willing to forgo a long life in poor health in exchange for a shorter life of better quality. The length of life they are willing to exchange indicates the relative preferences for different health states and this in turn can be used to weight QALYs (Clarke et al., 2010). If a health state is so bad, then it may be that people are willing to trade their whole life because it is seen as worse than death, however, the time trade-off method in this case cannot distinguish between different ‘worse than death’ health states (Devlin et al. 2011).

2.3.2.4 Contingent valuation

Contingent valuation is a trade-off based method used to provide monetary values for use in a CBA (Boxall et al., 1996). In a contingent valuation study, individuals are directly asked their WTP or willingness-to-accept (WTA) compensation for a particular event or scenario. WTP is an approximation of the compensating variation required to maintain each individual at the same ‘utility’ level, before and after the prospective change (Diener et al., 1998). WTA is an estimate of the equivalent variation, the amount an individual requires to achieve the utility level if a change had occurred (Diener et al., 1998). The WTP or WTA valuations can then be averaged from a sample of people and compared with total costs to see whether a good, service or change to provision is worth pursuing (O’Brien & Gafni, 1996).

Contingent valuation has been used extensively within many areas of economics (Venkatachalam, 2004; Klose, 1999). For example, in environmental economics, this method was used in a valuation of the compensation to be paid by Exxon after an oil leak in Alaska (Carson, 1994). Within health economics, contingent valuation has been used for estimating WTP for a range of goods and services. For example, the method was used to estimate values for mortality and morbidity risk reduction through WTP for preventative medication (Nielsen et al., 2012). Although there have been many studies published, contingent valuation has yet to be used in UK healthcare decision-making (Smith & Sach, 2010).
There are many criticisms of the method. Contingent valuation is often used to value unfamiliar goods (when revealed preference methods could not be used) but this means respondents may find it difficult to price something, particularly if it is health-related (Ryan et al. 2008). The contingent valuation survey approach is open to many known sources of bias including zero-bids (where respondents state ‘zero’ when they actually have a positive WTP), protest-bids (where respondents give artificially high values to influence decisions), anchoring effects (when respondents are influenced by their previous responses or suggested responses), ordering effects (where valuations differ depending on the order of questions) and payment vehicle effects (where certain terminology, for example using ‘tax’ or ‘subsidy’ influences WTP values) (Hanley et al., 2001; Holmes & Adamowicz, 2003).

Given these problems with using the contingent valuation method, methods such as conjoint analysis, DCEs and best-worst scaling, have been suggested as viable alternatives and are increasingly used in the healthcare context.

2.3.2.5.1 Attribute trade-off-based stated preference methods

Conjoint analysis, DCEs and best-worst scaling are three types of attribute trade-off-based stated preference methods. The methods all present a series of scenarios and ask respondents to state (through choosing, rating or ranking) the preferred scenario that describes a good or service using pre-defined attributes and levels. These methods are used to value individual attributes that make up goods or services and allow estimation of the trade-offs between different attributes. Trade-offs are defined as a balance of opportunity costs which maximise utility. Attribute based methods are grounded in Lancaster's Theory (1966) (see also section 2.4.1), which suggests consumers value particular features of goods or services rather than the product as a whole.

2.3.2.5.2 Conjoint analysis

Conjoint analysis (CJA) is a term used inconsistently in the published literature. Both the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and Sawtooth survey software (Sawtooth, 2012) refer to DCEs as CJA (Bridges et al. 2011; Louviere et al. 2010). However, this thesis distinguishes between the approaches in line with the recommendations of Louviere et al. (2010), and defines CJA specifically as a method which requires respondents to rate or rank hypothetical alternatives which are described by attributes of varying levels (Ryan et al., 2008).
Under ranking CJA, respondents are asked to order the alternatives to represent their preferences for each alternative (Boyle et al., 2001). In rating CJA, respondents are asked to consider not just the order of their preferences but also the strength of their preference (Boyle et al., 2001). Inclusion of a cost attribute allows analysis of the exchange of money for different levels of other attributes and thus we can derive their value and WTP (Johnson et al., 2011). CJA has been used for identifying preference weights to generate QALYs. It can therefore be thought of as both welfarist and extra-welfarist in its ability to inform both CBA and CUA (Flynn, 2010).

The development of specialised Sawtooth software (Sawtooth, 2012) has stimulated the use of adaptive conjoint analysis (ACA) (Fraenkel, 2010). In ACA, the choice sets are presented to respondents as either a rating or ranking task that are then updated based on their previous choices. The ACA approach is used as an attempt to improve the efficiency of the experimental design (Fraenkel et al., 2001). However, ACA surveys have been criticised for their time-burden (with surveys taking longer to complete than traditional designs) which could result in lower completion rates (Cunningham et al., 2010).

CJA is neither recommended, nor used, in the UK decision-making processes. The CJA ranking or rating task is not necessarily representative of a typical decision-making process. When choosing a good in a supermarket, individuals rarely rank or rate the goods. Louviere et al. (2010) further criticise CJA because it does not align with neoclassical economic theory that relies on a discrete choice rather than ranking or rating. This means that the analysis is not underpinned by economic theory and so the models used to explain uncertainty in typical economic theory are not applicable. In addition, because the data elicited are not representative of actual decision-making behaviour, it is not appropriate to compare the stated preference data with revealed preference data (Louviere et al., 2010). Louviere et al are quite clear when they state: “Conjoint analysis should be seen for what it really is, namely a purely descriptive way to fit a statistical model to a set of observed ranking or rating data with no ability to inform questions about how consumer behaviour is likely to change in response to changes in the choice context” (p. 69).

2.3.2.5.3 Discrete choice experiments (DCEs)

DCEs ask respondents to make a choice, usually pair-wise, to indicate their preference between a discrete set of two or more alternatives describing goods or services using attributes and levels (Ryan et al., 2008). Representing a choice task as used in DCEs is arguably more natural because they more closely represent decisions made by humans in everyday life. As DCEs involve respondents completing a relatively simple task, choosing
only one option, but the elicited data will be ‘weakly ordered’ (Louviere et al., 2010). ‘Weakly ordered data’ means information about the strength of preference for each individual attribute is not completely attained. However, there is a balance between using methods which provide stronger ordering data, such as CJA, and task complexity.

If a cost attribute (or price proxy) can be included in a DCE then it is possible to estimate a monetary valuation for the product being valued and therefore provide benefit measures for use in a CBA (Ryan & Gerard, 2003b). Furthermore, the relative importance of the different attributes included in the DCE can also be quantified. If health states are used as the attributes in the DCE, the method can also be used to generate preference weightings for the different components of a multi-attribute health status measure and the preference weights used to create utility values subsequently used in a CUA (Brazier et al., 2012; Stolk et al., 2010; Flynn, 2010).

Within the discipline of health economics, applications of DCEs have gone beyond providing data for economic evaluations and now answer research questions in a variety of contexts and settings (Clark et al., 2014; de Bekker-Grob et al., 2012). Two systematic reviews of published DCEs identified that the method had been used to address a variety of questions such as: estimating preference weights (utility); valuing health outcomes; measuring patient experiences; investigating the trade-off between health outcomes and patient experiences; investigating job-choices; and developing priority setting frameworks (de Bekker-Grob et al., 2012; Clark et al., 2014).

2.3.2.5.4 Best-worst scaling

A relatively new valuation method, often seen as an extension to DCEs, is best-worst scaling (BWS). There are three distinct cases of BWS methods (case one, case two, and case three). In case one, also known as object case, respondents choose the attribute that is the best and the attribute that is the worst (Louviere & Flynn, 2010). In case two, profile case, respondents choose the best attribute and the worst attributes in an alternative (Ratcliffe et al., 2012). The third case, multi-profile case, is the most similar BWS task to a traditional DCE as respondents choose the alternative that is the best and the alternative that is the worst from a choice set (Marti, 2012).

BWS is argued to be advantageous compared with DCEs because the choices made reveal more information about the strength of people’s preferences through fewer choice sets which could, in turn, reduce the response error (Xie et al., 2014). However, studies comparing BWS and other methods have found it produces similar but slightly less reliable estimates (Ratcliffe et al., 2011). A recent study by Whitty et al. (2014) used qualitative
interviews to compare the validity and acceptability of DCEs and BWS, and found more support (in terms of consistency and trading behaviour) for the traditional DCE task.

2.4 Theoretical foundation of DCEs
DCEs and BWS share a solid theoretical foundation that underpins the implementation of the methods and interpretation of the results. In contrast, CJA does not share the same theoretical basis. Rating or ranking experiments are not considered a useful approach for eliciting preferences, and therefore are not discussed further in this thesis. BWS was an emerging method and was not perceived to be a relevant method for eliciting benefit-risk trade-offs when this thesis was conceptualised. Traditional DCEs have been used more often than BWS and are a more established method (Clark et al., 2014).

Furthermore, given evidence suggesting that DCEs are less burdensome and possibly produce more reliable results, a traditional ‘choose the most preferred’ choice experiment was selected as the focus for this PhD. For this reason the remainder of the thesis focusses on the design, analysis and application of DCEs. DCEs are based on two important theories: Lancaster’s Theory and RUT described in section 2.4.1 and 2.4.2, respectively.

2.4.1 Lancaster’s theory
Prior to the 1960s, it was traditionally thought that consumers valued goods and services as whole ‘offerings’ and decisions were made on psychological taste for the good or service rather than on their distinguishing features (Debreu, 1960; Johnson, 1958). In the 1960s, this traditional economic theory was challenged and new theories developed to try and explain choice behaviour; suggesting goods or services are made up of attributes or characteristics (Lancaster, 1966; Alcaly & Klevorick, 1970).

All attribute-based valuation methods, including DCEs, are based on theories developed by Lancaster in his 1966 paper: A New Approach to Consumer Theory. This paper queried the traditional theories of consumerism that “goods are goods” (p.132, Lancaster 1966) and suggested three hypotheses: 1) consumers value the attributes that a good possesses, rather than the good itself; 2) goods are made up of many attributes which are not necessarily unique to that good; and 3) two goods together may possess different attributes from those when they are separate. DCEs incorporate these three hypotheses through the development of alternative profiles, choice sets and interactions between attributes, respectively.

A key inference to be made from Lancaster’s paper is that goods are distinguishable because of their attributes, and possess inherent differences on which a consumer chooses. This theory allows an objective explanation of why some things are good substitutes and
why others are complements, rather than relying on the assumption of ‘intrinsic properties’ that traditional consumer theory depended upon (Johnson, 1958). Another key inference that can be drawn from Lancaster’s theory is that the values placed on each characteristic can be summed to estimate the value of the good or service as a whole (Ryan, 2004).

2.4.2 Random utility theory

The way individuals choose between the alternatives presented in a DCE can be explained using probabilistic choice theory. Two hypotheses can explain choice behaviour: 1) decisions are random and utility is deterministic; and 2) decisions are deterministic and utility is random (and ‘actual’ behaviour cannot be modelled). In the first hypothesis, the individual can be assumed to choose on impulse as influenced by psychological factors (Tversky, 1972). In contrast, the second hypothesis regards individuals as utility maximisers, but there is a random component to this maximisation (Thurstone, 1927). The second hypothesis, that suggests decisions are not made randomly but rather utility has a random component, has been explored most in the context of DCEs through random utility models to explain behaviour and provides the model for this thesis (Ryan et al., 2008).

The random utility models used to explain the uncertainty around predicting consumer and respondent choices are underpinned by RUT. RUT was originally investigated by Thurstone (1927) who looked at the derivation of satisfaction, or utility, through a ‘Law of Comparative Judgement’ with a psychological perspective. The theory was developed substantially in the 1970s with econometric input from the 2000 Nobel Prize Winner, Daniel McFadden (McFadden, 1974, 1986).

RUT provides a deterministic-decision framework which is not trying to explain irrational behaviour, but model the researcher’s lack of information. The lack of information results in an error which could be due to measurement errors or latent attributes that influence choice or heterogeneity in preferences (unobserved differences in taste). Therefore, the psychological factors which influence choice are incorporated into this random component of utility (Louviere et al., 2010).

RUT is based on the simple axiom that it is not possible to observe the ‘actual’ utility function, however, it is possible to infer what is affecting utility from deterministic decisions being made (Louviere et al., 2010). Utility \( U_{in} \) is said to be a latent construct that people hold in their mind (hence it is unobservable), and this construct is made of both systematic \( V_{in} \) and random \( \varepsilon_{in} \) components (Louviere et al., 2000). In equation 2.1, \( U \) is the unobservable utility of alternative \( i \) for individual \( n \), \( V \) is the deterministic utility and \( \varepsilon \) is a random unobservable component:
\[ U_{in} = V_{in} + \varepsilon_{in} \]  \[2.1\]

In a DCE, \( V \) is constructed of: the attributes and levels in the alternative scenarios presented; and covariates to explain individuals (such as socioeconomic differences). The random component reflects all unobservable factors influencing decision-making behaviour. This random element to behaviour and decision-making means it is not possible to accurately predict the alternative the respondent will choose and therefore choices are modelled as probabilities.

RUT assumes individuals seek to maximise utility, \( U \), but as their actual utility function is unobservable, a probabilistic utility function should be used to estimate their choices. This probabilistic utility function estimates the likelihood of the individual choosing an alternative out of a set of feasible alternatives. Equation 2.2 illustrates this concept and defines the probability of choosing alternative \( i \) over another alternative \( j \) as:

\[
P_i = \text{Prob}(U_i > U_j) = \text{Prob}(V_i + \varepsilon_i > V_j + \varepsilon_j) = \text{Prob}(V_i - V_j > \varepsilon_i - \varepsilon_j) \]  \[2.2\]

The first part of equation 2.2 explains that the probability that the utility from alternative \( i \) is greater than alternative \( j \) is the same as the probability of the deterministic utility of alternative \( i \), given random errors, is greater than the deterministic utility of alternative \( j \). This implies that if there is a high probability of choosing alternative \( i \) there is considerably more utility offered upon its consumption than alternative \( j \).

If, in a particular set of alternatives, \( i \) has more desirable levels in its attributes, then the probability of choosing alternative \( i \), \( P_i \), tends towards one because the difference in deterministic utility between alternatives \( i \) and \( j \) increases. If the utility difference was only small, then the probability would tend towards 0.5 as it is more difficult for the individual to choose between the alternatives, and therefore the decision becomes more random; like tossing a coin. Estimates of these probabilities can be calculated from data elicited from a DCE by looking at the choices made by the respondents in each set as the proportion of respondents choosing alternative \( i \) represents the probability of an individual choosing \( i \).

Popular discrete choice models include the conditional logit, multinomial logit (MNL) and mixed logit (MXL) (de Bekker-Grob et al., 2012; Clark et al., 2014). There are no agreed gold standards for the design, analysis or interpretation of a DCE. However, there are recommendations published by organisations and experts in the field (Bridges et al., 2011; Lancsar & Louviere, 2008; Ryan & Gerard, 2003a). These guidelines were used to inform the DCE conducted as part of this thesis (see Chapter Three).
2.5 Risk in healthcare valuation

Within healthcare, the number of studies using DCEs has been gradually increasing since 1990, and the number containing risk as an attribute has also increased (de Bekker-Grob et al., 2012; Ryan & Gerard, 2003b; Clark et al., 2014; Harrison et al., 2014).

2.5.1 Definition of risk

The British Medical Association (BMA) defines risk as “the probability that an adverse event will occur, analogous to the likelihood, the odds or chance” (p.6, Mansfield et al. 2012). In the simplest definition, risk can be thought of as concept linking the probability of an outcome, usually negative, occurring and the severity of that outcome (Mansfield et al., 2012). A risk may be categorised as ‘high’ because of the magnitude of its likeliness or because it has a very unpleasant outcome (Lipkus, 2007). Similarly, a negative risk (probability of something bad happening) can be re-framed to be a positive risk (the probability of something good). For example, the risk of mortality (a negative frame) can be translated into a positively-framed chance of survival (Edwards et al., 2001). Although often used interchangeably, risk is distinctly different from uncertainty. For the purpose of this thesis, a definition by Mishel (1990) is taken which describes uncertainty as an unknown, or unquantifiable, probability associated with an event occurring.

Risk is made up of multiple components. When considering risk, individuals may take into account: the severity of the possible outcome and its irreversibility; the baseline level of risk that is typical for everyone; the duration of risk exposure and whether they will return to a baseline level; event time; the certainty surrounding the risk; the objectivity of the risk and their preformed perceptions about their likelihood; and risk latency with regards to whether the risk will occur now or in the future (Hammit & Graham 1999).

Another key component is whether the risk is viewed as being voluntary or imposed on an individual (Edwards et al., 2002). Risks can be voluntary when, for example, there is an established link that being overweight can create an increased risk of diabetes and yet individuals do not lose weight (Abdullah et al., 2010). Risks can also be imposed on an individual over which they have no control, for example, exposure to polluted water and the risk of parasitic disease (Wu et al., 1999). Other risks are combinations of voluntary and imposed risks, such as a genetic predisposition to a condition which may be increased by certain lifestyle factors (Criswell et al., 2006). The different components of risk are defined in Table 2.2.
Table 2.2 Components of risk

<table>
<thead>
<tr>
<th>Component</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severity of outcome</td>
<td>The significance of the hazard or event.</td>
</tr>
<tr>
<td>Irreversibility of outcome</td>
<td>The opportunity to undo the event and return to a previous health state.</td>
</tr>
<tr>
<td>Baseline level of risk to everyone</td>
<td>The risk elevation for an individual.</td>
</tr>
<tr>
<td>Duration of exposure to risk</td>
<td>The length of time an individual is exposed to a risk which could be temporary or permanent.</td>
</tr>
<tr>
<td>Event time horizon</td>
<td>The time until the hazard will occur; this could be immediate or long-term.</td>
</tr>
<tr>
<td>Certainty of risk estimates</td>
<td>The risk could be calculable as a point estimate or defined with confidence intervals.</td>
</tr>
<tr>
<td>Objectivity of risk</td>
<td>The risk statistic may be applicable to a whole population or specific to an individual.</td>
</tr>
<tr>
<td>Perceived risk and experience</td>
<td>The risk may be influenced by an individual’s experience of the hazard.</td>
</tr>
<tr>
<td>Risk latency</td>
<td>The risk may not be immediate but something that becomes elevated in the long-term.</td>
</tr>
<tr>
<td>Imposed and voluntary risks</td>
<td>An individual may or may not have control over their risk for a particular hazard.</td>
</tr>
</tbody>
</table>

Source: Hammitt & Graham (1999)

2.5.2 Risk in healthcare

Risk is a ubiquitous component of health and healthcare delivery in the form of iatrogenic effects. In addition, lifestyle choices such as overeating or smoking can cause a risk resulting in a negative effect on health. A balance between benefits and risks when making decisions about health can occur at various stages of care. Table 2.3 provides examples of different risks an individual may face relating to their health or healthcare choices although this list is not exhaustive.

Table 2.3: Examples of risk in types of healthcare

<table>
<thead>
<tr>
<th>Type of healthcare</th>
<th>Preventative</th>
<th>Prognosis</th>
<th>Diagnosis</th>
<th>Treatment</th>
<th>Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example risk</td>
<td>Risks associated with lifestyle</td>
<td>Risk of mortality</td>
<td>Risk of misdiagnosis</td>
<td>Risk of an adverse event</td>
<td>Risk of a false negative/positive</td>
</tr>
</tbody>
</table>

The potential for interactions between Tables 2.2 and 2.3 show that risk is a complicated and multifaceted concept and constitutes more than a statistical statement of a probability.

2.5.3 Risk in a DCE

As a ubiquitous part of healthcare delivery, there are many situations when individuals can be faced with different types of risk in the process of their treatment and for this reason risk is a commonly occurring attribute in DCEs (Ryan & Gerard, 2003b; de Bekker-Grob
et al., 2012). In line with previous reviews, a recent systematic review by Harrison et al. (2014) showed an exponential year-on-year increase in the number of DCEs containing risk attributes since the 1990s. One explanation for this increased use is that incorporating risk in a DCE will allow estimation of the maximum acceptable risk for a benefit and analysis of preference proportionality (Johnson et al., 2009). If a price proxy is also included, it allows information on the WTP for a risk reduction and an alternative valuation of a statistical life (Telser & Zweifel, 2007).

Recently, there have also been suggestions that quantitative estimates of benefit-risk preferences could be used to support regulatory decision-making as a complement to the clinical evidence of interventions (Hauber et al., 2013). With this new potential to inform drug approval decisions it is important that the methods used to elicit people’s benefit-risk thresholds are accurate and robust.

Whilst valuation of benefit-risk trade-offs are useful, some studies that presented risk in stated preference surveys have found it has an insignificant effect on preferences or it produces values which do not align with the value of statistical life estimates (Corso et al. 2001; Watson et al. 2009). Corso et al. (2001) proposed that benefit-risk valuations from stated preference methods may appear to be inaccurate because of the following hypotheses:

1) Risk estimates have already been formed by respondents who have their own perceived risk;
2) Economic theories are not compatible with real human behaviour and assumptions, such as linear responses to risk, falsely indicate ‘incorrect’ answers;
3) Differences in risk magnitude are not being accurately communicated.

The underlying economic theories (described in section 2.4) supporting the use of DCEs as a valuation method require attributes to be interpretable by respondents. The presence of any of the hypotheses from Corso et al. (2001) could introduce bias and invalidate any valuations derived from a DCE.

Hypotheses 1) and 2) could also be due to an increase in the cognitive burden of a DCE which may force respondents to employ simplifying heuristics ignoring the risk attributes. If this occurs, then no information about the trade-offs of risk can be made (Bridges et al., 2011a). Additional support for hypothesis 3) has been confirmed by other authors, for example, Watson et al. (2009) included risk of pelvic inflammatory disease from Chlamydia in their DCE and hypothesise that: “the insignificance of the risk of pelvic inflammatory disease may reflect the difficulties that respondents had in understanding this attribute” (p. 623). A systematic review by de Bekker-Grob et al. (2012) stated:
“Studies have continued to include risk as an attribute. [Previous reviews] noted the difficulties individuals have understanding risk, and they commented on health economists giving little consideration to explaining the risk attribute to respondents. There appears to be little progress here... This is clearly still an important area for future research.” (p.150).

In a more recent systematic review of risk in DCEs, Harrison et al. (2014) found a lack of training, information and descriptions with fewer than 20% of the studies reviewed containing risk communication aids. The authors also noted that those designs that did contain risk communication aids were published only recently. Also, evidence from other valuation methods such as the standard gamble and contingent valuation, has shown that risks are often misinterpreted by respondents when no aids are provided (Sharma et al., 2002; Corso et al., 2001).

It may be appropriate to provide risk communication aids to respondents to enhance the respondents’ understanding of the DCE and their ability to make comparisons and trade-offs. Possible risk communication formats which may aid respondents include comparative graphs, risk grids or risk ladders (Schapira et al., 2001). However, there are many different risk communication formats and, as yet, there is no evidence of their relative effectiveness for trade-off decisions. Exploration of the different risk communication formats which could be used in to aid respondents understanding of attributes in a DCE is expanded on in a structured literature review presented in Appendix 3.1 and described in section 3.2 of Chapter Three.

2.5.4 Decisions with risk

Every day, people are confronted with decisions to make about risky behaviour. Many decisions are made based on gut feeling rather than on the interpretation of a mathematical probability, and this if often because no explicit statements about the risk exist (Slovic et al., 2005). For example, choices about what to eat or a mode of transport to travel on are often based on ‘feeling’ or intuition rather than ‘analysis’ or logical reasoning (Slovic et al., 2005).

Non-perfect reason based on feelings of fear helped humans evolve a fight or flight reaction to a hazard (Renn, 2006). In risky situations, humans have had to make quick decisions and judgements based on instinct or intuition rather than hesitating to assess the situation and take into account probabilities. Whilst this has worked well for human survival, it has resulted in a brain inherently poorly equipped to reason with factual risks (Kahneman, 2012). This cognitive shortcut is described by Slovic et al. (2005) as the
‘affect heuristic’. As a consequence of the ‘affect heuristic’, when risk information is presented, humans’ evolutionary traits prove an obstacle to effective communication.

There are, however, occasions where individuals need to weigh-up situations with numbers and facts of risk likelihood. In healthcare, there have been increasing calls for shared decision-making with individuals (patients, clinicians) deciding treatment based on an assessment of the potential consequence and their likelihood (Charles et al., 1997; Stiggelbout et al., 2012). Choices about the benefits and risks of healthcare treatment should go beyond that of gut feeling and minimising the affect heuristic to communicate risk effectively is an essential part of accurately eliciting people’s preferences.

2.5.4.1 Expected utility theory

When choosing between alternatives that are risky or uncertain, an individual must evaluate the presented prospects under risk or ambiguity, respectively. In behavioural economics, decision-making under risk or ambiguity is explained through a number of theories. One well explored theory is Expected Utility Theory (EUT). EUT suggests individuals will weight the possible outcomes by their likelihood (Von Neumann & Morgenstern, 1947). Whilst it is assumed that an individual will still seek to maximise their utility, it is also assumed that they will consistently choose the alternative which is in line with their risk preferences and reject options which involve too much risk for too little reward (Rabin, 2000).

In addition to EUT, there are a number of generalisations which could also explain respondents’ behaviour. Hey & Orme (1994) looked at experimentally generated preference data to see how the different generalisations (such as prospective reference theory, weighted utility theory, regret theory and disappointment aversion theory) and different restrictions breaking neutrality (such as absolute risk aversion) fitted with their collected data. Hey & Orme (1994) showed that, in the stated preference data, expected utility held and, with additional structure of the error term, it could predict individual choice. However, other studies have found EUT was frequently violated, with preferences more in line with prospect theory and biases induced from loss aversion (Rabin, 2000; Avineri & Prashker, 2004). RUT overcomes some of the shortcomings of EUT and is the standard method for the analysis of choice data as it allows for unobserved, random, influences on choice and will therefore be used to analyse choices in the empirical chapters of the thesis (Lloyd, 2003).
2.5.4.2 Risk literacy

There is a substantial evidence base (see Lipkus et al. 2001; Gigerenzer et al. 2007) which shows a general lack of understanding about risk across all demographics, possibly related to the psychological reasons described in section 2.5.4. Effective risk communication is also susceptible to the challenges of any numerical or scientific information. When asked which risk was biggest out of: 1 in 10; 1 in 100; or 1 in 1,000, 28% (n=282) of Germans and 25% (n=249) of Americans provided an incorrect answer (Galesic & Garcia-Retamero, 2010). The authors of this study suggested that because risks with the largest magnitude were described with the smallest denominator frequencies, the subjects became confused.

Similarly, Denes-Raj & Epstein (1994) asked individuals to choose between two bowls: 1) a bowl with nine white beans and one red; and 2) a bowl with 100 beans with nine or fewer being red. The individuals who took part won a prize of $7 if they picked one red bean. Despite knowing that probability of bowl two was against them, the individuals were observed to indicate that they felt they had a better chance when there were more red beans available. This ‘ratio bias’ has been replicated in other studies and emphasises the challenges of choosing framing affects using different numerators and denominators. There are multiple examples of biases and framing effects in the risk communication literature; the term ‘collective statistical illiteracy’ has been used to describe these challenges (Gaissmaier & Gigerenzer, 2008; Gigerenzer et al., 2007).

2.5.5 Communicating risk information

Communicating information on risk introduces the concept of framing affects and how the choice of risk information and its presentation may affect the interpretation of the probability statistic.

2.5.5.1 Absolute and relative risk

A popular example of miscommunicated risk is the 1995 ‘pill scare’ where a new generation of the contraceptive pill was found to increase women’s risk of thrombosis by 100% (Committee on Safety of Medicines, 1995). The subsequent newspaper headlines failed to explain an increased relative risk of 100% was an increase in absolute risk from 0.014% to 0.028%. The news coverage and misleading risk information resulted in an estimated 13,000 additional abortions at a cost of around £46 million to the NHS (Furedi, 1999). The example of the ‘pill scare’ illustrated the importance of using absolute risk information to communicate risk.

A key step in transparency is providing individuals with enough information to make their own choices (not necessarily the ‘right’ answer). A lack of information means people make
their own inferences of the situation using subjective value judgements. For this reason, it is generally accepted that risk should be communicated as an absolute value rather than relative risk.

2.5.5.2 Sufficient information

In addition to presenting the appropriate statistic, the risk information should be contextualised in a transparent way. A study by Gigerenzer et al. (2005) surveyed a typical weather forecast to understand how the information was interpreted by members of the public. When asked what a “30% chance of rain tomorrow” meant, most respondents interpreted this to mean that there would be rain 30% of the time or in 30% of the area. The correct response, that it will rain 30% of the days like tomorrow, was the least chosen alternative in the multiple choice question. Gigerenzer et al. (2005) hypothesise that this could be due to ambiguity of such statements and stress the importance of re-iterating risk information using clear and simple terminology and, where possible, examples.

2.5.5.3 Positive and negative frames

Another challenge is the framing of the risk. A positive ‘survival rate’ can be interpreted differently to a negative ‘mortality rate’. For example, Tversky & Kahneman (1981) asked individuals:

Imagine that you face the following pair of concurrent decisions. First examine both decisions, then indicate the options you prefer.

i) Choose between:
   A. A sure gain of $240
   B. 25% chance to gain $1000 and 75% chance to gain nothing

ii) Choose between:
   C. A sure loss of $750
   D. 75% chance to lose $1000 and 25% chance to lose nothing

Most of the study respondents chose the risk averse option of A in decision i), and the risk seeking option of D in decision ii). When risk was framed positively, it was seen that there was a potential gain to be made and the behaviour was more risk averse. However, when risk was framed negatively, the potential for a loss induced more risk-seeking behaviour. These behaviours are systematic and consistently present in choices about financial or health risks (Fishburn & Kochenberger, 1979; McNeil et al., 1982; Kahneman & Tversky, 1979).
In healthcare, the appropriate framing depends on the objectives of the communication as framing can influence patients’ decisions for healthcare (Haward et al., 2008). In many areas of public health, risk communication is intended to change actions or behaviour (Rothman & Salovey, 1997). However, risk communication in a DCE context is not about changing behaviour but instead about effectively informing respondents so they can make a decision that is right for them and reflects their preferences.

2.5.5.4 Risk communication formats

One of the key challenges of communicating risk is to encourage risk literacy without controlling or manipulating people’s feelings, preferences or motives. To help the objective communication of risk, a number of risk communication aids have been created. In addition to different numerical expressions of the information (percentages, frequencies, decimals, odds ratios), visual stimulants such as graphs or pictures have been developed to help people’s interpretation (Peters et al., 2011; Lipkus & Hollands, 1999). As there are many different ways of presenting the probabilistic information to people, the overwhelming literature needed to be approached using a structured and specific focus to identify possible risk communication formats for use in a DCE.

2.6 Conclusion

This chapter has described why valuations using stated preference methods of current and new interventions are useful in the UK healthcare system. The chapter also identified DCEs as an appropriate method for the valuation of healthcare goods and services by comparing its strengths and limitations with alternatives such as contingent valuation and CJA. Evidence of the challenges of using value attributes, particularly risk, in a DCE were also identified by the literature review.

The chapter concluded that if DCEs are to be used to elicit people’s preferences for benefit-risk trade-offs, it is imperative that the information is communicated effectively so respondents can make informed choices which accurately reflect their preferences. The following chapter, Chapter Three, describes the development of the DCE used in the empirical chapters of this thesis.
Chapter Three
Valuing risks and benefits: a case study of population-based screening for breast cancer

3.1 Introduction
The aim of this chapter is to report the development of a DCE to value respondents’ preferences for benefit-risk trade-offs. The chapter addresses two main objectives: 1) to identify a suitable case study; and 2) to design a DCE. The chapter is presented in three sections. The first section (starting at section 3.2) describes the identification of risk communication formats to be investigated in the DCE. Section 3.3 starts by describing key criteria for selecting a case study to value benefit-risk trade-offs, and how population-based screening for breast cancer was selected. The second section (starting at section 3.4) explains the clinical background of breast cancer, screening programmes, the type of benefit-risk trade-offs involved in participation and the implications of generating quantified preference data for decision-makers seeking to allocate resources efficiently. The third section (starting at section 3.5) describes the process for generating the DCE design including possible attributes and levels, experimental properties and pilot studies. The chapter finishes with a description of the final DCE to be used in the empirical studies described in Chapters Four, Five and Six.

3.2 Risk communication formats
Prior to the selection of a case study, two risk communication formats were identified. These formats were found through a structured review which aimed to systematically identify different formats which can be used to communicate risk information to DCE respondents. If risk is communicated effectively, then it will assist the respondents in making informed choices based on their preferences and thus improve the quality of the data collected. A complete description of the review can be found in Appendix 3.1. The review was also published as a subsection in a paper by Harrison et al. (2014) which can be found in Appendix 3.3.

A systematic search of five electronic databases (Medline, Embase, Web of Science, PsychINFO and EconLit) was conducted in April 2013. The search checked the titles of articles in each database for the terms risk and either communication or format, and the whole abstract for topic terms (aid, presentation, display). The exact strategy included the following terms: Title=(risk*) AND Title=(communicat* or format*) AND Topic=(aid* or present* or display*).
The search of all databases revealed 1,207 possible hits. After the removal of non-peer-reviewed material and duplicates, 390 titles and abstracts were reviewed. In total, 215 full articles were retrieved for further review and 99 papers were included in the final review. Of the included studies, 65 were empirical tests of risk communication formats, 21 were overviews of the risk literature with no reported search strategy, nine were systematic reviews with a structured search strategy, and four looked at presenting uncertainty (for example, the use of confidence intervals as opposed to point estimates). A list of all included studies can be found in Appendix 3.2.

The results of this review uncovered a fine line between providing sufficient information for respondents to understand the risk and overwhelming them with too much. Training materials for DCE respondents should therefore be presented at a level that is accessible to all but should not be so elementary that respondents will ignore it. A prescriptive approach to the DCE, with extensive piloting and revisions, should therefore be conducted to ensure materials are appropriate for respondents.

The identified empirical studies also highlighted the importance of assessing numeracy skills on a respondent’s ability to interpret probabilistic information. Preformed ideas about perceived risk based on experience were also identified as a challenge to effective communication of risk information. In understanding how respondents interpret risk in a DCE, numeracy, experience and perceptions could be useful covariates to collect in supplementary questions in a DCE study.

The review identified icon arrays as the risk communication format with the most empirical support which would make it an appropriate comparator to current practice in DCE literature which are percentages only (Harrison et al., 2014). However, no empirical study investigated the use of icon arrays in a trade-off task so it is unclear whether they would be superior in a DCE context. The next sections of this chapter describe the design of a DCE to incorporate the selected risk communication formats (icon arrays and percentages; and percentages only) identified by the review.

3.3 Selecting a case study

This PhD used a case study to quantify benefit-risk trade-offs and investigate the impact of how risk is communicated in the DCE. A list of case study selection criteria were developed based on practical considerations and literature identified in Chapter Two. The criteria provided a pragmatic approach to selection and aimed to ensure the chosen case study would be feasible and a suitable context in which to investigate the effect of different risk communication formats in a DCE.
3.3.1 Case study selection criteria

Eight criteria were used to select a case study:

Criterion 1: The good or service described in the DCE must represent at least one clear example of the different types of risk. As risk manifests itself in healthcare through multiple means (see Tables 2.2 and 2.3), a case study for this PhD thesis should incorporate at least one of these risks in order to investigate the effect of different communication formats.

Criterion 2: The disease area should have significant burden, incidence and prevalence rates as proxies for interest to decision-makers in primary care. Primary care is the first point of access and often the only point of care for many patients. This PhD received funding from the National Institute for Health Research (NIHR) School for Primary Care Research (SPCR). Therefore, the outputs of the research must produce results applicable to primary care and be applicable to a large number of current and potential patients.

Criterion 3: Consumption of the good or services described by the DCE should result in a quantifiable risk. As discussed in Chapter Two, risk and uncertainty are distinctly different concepts with different underlying theories, and with the latter representing an unknown probability of the hazard occurring. The focus of this PhD thesis is the best way of communicating risk attributes which implies that the probabilities must be identifiable and quantifiable.

Criterion 4: Opting-out of the choice set should be a realistic and plausible alternative. If respondents are particularly risk averse it would be advantageous to have a ‘no treatment’ option for them to choose. Without the ability to opt-out, the estimated valuation may be upwardly biased, implying the DCE respondents are more tolerant of risk than they truly are. In order to add a realistic opt-out, the risk probabilities must be available for current care, the status-quo or no intervention.

Criterion 5: The case study should contribute to the DCE literature with an original application. As there are a growing number of applied DCEs using a variety of healthcare goods and services, it would be preferable to conduct original research rather than duplicate an existing study.

Criterion 6: Access to experts to verify the DCE is appropriate in its use and description of the attributes and levels. There are practical considerations in conducting a high quality DCE in the time span of a PhD project. If the case study could tie in with other areas of
research where patient and expert involvement is already established, this could improve the rate of study progression.

Criterion 7: Stated preference methods should be appropriate and the outcomes of the DCE should be of interest to decision-makers. Preferably there would not be revealed preference data available so the results of a stated preference study are of use. This could be because the intervention is not currently provided by the NHS or because there is no market data due to the reasons discussed in Chapter Two, section 2.3.1.

Criterion 8: The case study should ideally be familiar, in order to minimise respondents’ confusion about the clinical context. This criterion aims to minimise the other sources of cognitive burden in a DCE task by ensuring the selected case study is not too complex. This means that respondents are likely to have better formed preferences and as a consequence the training materials prior to the choice sets can be briefer. A familiar topic may also keep the DCE respondents engaged in the task, generating higher quality data.

3.3.2 Identification and appraisal of case studies.

A number of suggested topics were discussed as potential case studies using the criteria described in section 3.3.1. These suggestions were largely driven by originality of the research and links to research on-going in the department which would allow for rapid access to experts, healthcare professionals and patients. Table 3.1 compares the possible case studies based on these criteria.
<table>
<thead>
<tr>
<th>Service</th>
<th>Case study</th>
<th>Prevalence</th>
<th>Available treatment</th>
<th>Primary care</th>
<th>Risk attribute</th>
<th>Life-style risk</th>
<th>Life-time risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic testing</td>
<td>Age-related macular degeneration (AMD)</td>
<td>Approximately 1.64% of UK population.</td>
<td>Yes (not for all types and not all on NHS)</td>
<td>Potentially results from a general practitioner (GP).</td>
<td>Risk of AMD with current lifestyle; risk of false negatives/positives.</td>
<td>Yes</td>
<td>6% for Women; 3% for males</td>
</tr>
<tr>
<td></td>
<td>Retinitis pigmentosa</td>
<td>Approximately 0.025% of UK population.</td>
<td>No</td>
<td>No current link.</td>
<td>Risk of developing retinitis pigmentosa; risk of false negatives/positives.</td>
<td>No</td>
<td>Unknown</td>
</tr>
<tr>
<td>Pharmacotherapy</td>
<td>Hypertension</td>
<td>3% &lt;40, 28% 40-79, 50% 80+ years old of UK population.</td>
<td>Yes – anti-hypertensive drugs</td>
<td>Yes</td>
<td>Risk of disease with/without drug; risk of adverse events (risk of stroke, heart failure, aneurysm); risk of side effects.</td>
<td>Yes</td>
<td>80-90%</td>
</tr>
<tr>
<td></td>
<td>Hypercholesterolemia</td>
<td>Approximately 3% of UK population.</td>
<td>Yes – statins</td>
<td>Yes</td>
<td>Risk of disease with/without drug; risk of adverse events; risk of side effects (risk of heart disease).</td>
<td>Yes</td>
<td>Risk of heart disease increased</td>
</tr>
<tr>
<td>Screening</td>
<td>Breast cancer</td>
<td>Most prevalent cancer in women.</td>
<td>Yes – chemo-preventative tamoxifen</td>
<td>GP prescribing, referral and aftercare. Some screening in primary care vans.</td>
<td>Risk of radiation induced cancer; risk of false negative; risk of false positive (over diagnosis); risk of mortality from breast cancer.</td>
<td>Complex</td>
<td>12.5% for women</td>
</tr>
</tbody>
</table>

1 O’Shea & Harvey (2003)
3 Cancer Research UK (2010)
4 Lane & Lip (2001)
5 Midlands Therapeutics Review & Advisory Committee (2008)
6 Women’s Eye Health Task Force (2003)
7 Department of Health and Human Services (2004)
8 Austin et al. (2004)
Three broad categories were identified as risk-prevalent areas of healthcare: genetic testing; pharmacotherapy; and screening. Two new genetic tests were identified for the inherited eye conditions: AMD and retinitis pigmentosa. The two conditions slowly lead to loss of sight and there are very limited treatments available (O’Shea & Harvey, 2003; Bundey & Crews, 1984). As a consequence of lack of treatment, understanding people’s preferences for these tests would introduce the need to value the role of information. However, the genetic test for AMD was so new that the risk of false negative or positives from the test was uncertain (with large confidence intervals) and therefore the case study did not satisfy criterion 3. The genetic test for retinitis pigmentosa was more developed, with known sensitivity and specificity, however, the prevalence of the condition was considerably lower and therefore the applicability of genetic testing would be for a small proportion of the population. A case study of a genetic test for retinitis pigmentosa would therefore fail criteria 2 and 8.

A case study of drugs for either hypertension (high blood pressure) or hypercholesterolemia (high cholesterol) would overcome the problems of genetic testing as these are prevalent disease areas with significant burdens. The disease area is also well researched with the risks of different treatments (hypertensive drugs or statins) well established through multiple long-term randomised controlled trials (RCTs) (LaRosa et al., 1999; ALLHAT, 2002; Lindholm et al., 2005). However, the potential risks are heavily influenced by lifestyle factors such as an individual’s smoking status, body mass index (BMI), engagement in physical activity, diet (potassium and salt intake), and alcohol consumption style (Boden-Albala & Sacco, 2000). Therefore calculating reasonable risk attributes for a generic DCE would be complicated and particularly vulnerable to subjective interpretation by respondents who may believe they are more or less at risk than average because of their behaviour. As a result, these two possibilities of pharmacotherapies for hypertension or hypercholesterolemia were dismissed.

Preventative screening is an intervention to detect illness at an early stage and aims to reduce the risk of the disease (Jepson et al., 2000). However, screening is rarely perfect and therefore consumption of a screening programme generates its own risks. For example, tests often have some degree of inaccuracy so there is often a chance of a false-positive or false-negative screening result. Understanding how these risks of participation in a programme must be balanced with the chance of detecting a disease is important for decision-makers trying to assess which services provide the most benefit. The UK’s Office for National Statistics (ONS) reports that one of the most prevalent illnesses is cancer, with different forms frequenting the top causes of mortality for both men and women (ONS,
Breast cancer is the most common cancer for women (Ferlay et al., 2010), and the leading cause of death for middle aged women (ONS, 2012).

As a consequence of its prevalence, studies surrounding breast screening are plentiful and there is a substantial evidence base comprising multiple randomised controlled trials and systematic reviews from across the world (Gøtzsche & Olsen, 2000; Kerlikowske et al., 1995; Ohuchi et al., 1995). Despite the UK offering screening, it is not mandatory and therefore an option of ‘No Screening’ is a plausible alternative in a breast screening DCE. The University of Manchester is also home to the largest independent cancer research organisation in the UK, with world experts in breast screening and breast cancer research working in the institution. Breast cancer and breast screening have also been the topics of long lasting public health campaign with targeted leaflets and media attention (Kamenova et al., 2014; Slaytor & Ward, 1998). The campaigns, combined with high prevalence, means that many people are familiar with the disease (McMenamin et al., 2005).

Breast screening was selected as a service to evaluate in a DCE for this thesis because it satisfied key criteria for a case study.
3.4 Screening for breast cancer
A national breast screening programme was identified as the most appropriate case study from the available options described in Table 3.1. The following sections describe breast cancer as a disease area, the need for a national breast screening programme and the debates surrounding the screening programme in the UK.

3.4.1 Background to breast screening
Breast cancer is a major health issue not just because of the prevalence and mortality risks, but also for financial reasons. Breast cancer costs the UK economy an estimated £1.5 billion pounds annually and the cost to the NHS of a single case of breast cancer is over £7,000 (Leal et al., 2012; Dolan et al., 1999). The high expense of general cancer care (an estimated £15 billion annually in the UK) means many Western countries now encourage participation in screening for common malignant diseases (von Karsa et al., 2008).

In England, the NHS currently invites all women between the ages of 50 and 70 years for screening using mammography (an X-ray of the breast) every three years as part of NHSBSP (Cancer Research UK, 2013). This programme is based on the premise that regular screening can identify tumours and ensure therapy is begun as soon as possible. In 2011, about 1.73 million eligible women were screened for breast cancer (excluding other referrals and those outside the current age range), an uptake rate of about 75% (The NHS Information Centre Screening and Immunisations, 2012). The NHSBSP is estimated to have an annual expense of £96 million (Public Health England, 2013).

3.4.2 Facts, figures and debates
Screening for breast cancer using mammography has been proven to detect cases of breast cancer earlier, and in 2010-2011 an estimated 14,725 undiagnosed cancers were detected by screening (The NHS Information Centre Screening and Immunisations, 2012). Women who participate in screening programmes have been shown to have improved mortality rates because of earlier interventions (Independent UK Panel on Breast Cancer Screening, 2012). In addition, women who participate in screening and receive a true-negative have the reassurance of knowing that they are cancer-free (Johnston et al., 1998).

In any X-ray, there is exposure to small amounts of radiation which result in a very small risk of radiation-induced cancers (Yaffe & Mainprize, 2011). In addition, as the mammogram produces an image which is interpreted by a radiographer, there is a chance of a false-negative and a cancer being missed. There is also a risk that the image will locate a true cancer but one which is so slow growing it would never have been harmful in the woman’s life-time (referred to as ‘over-diagnosis’) (Welch & Black, 2010).
In addition to the risk of over-diagnosis, women can be recalled back to the screening centre for biopsies and further unnecessary tests as a result of a false-positive (Hofvind et al., 2004). These unnecessary follow-ups are not only painful, but an incorrect test result can cause undue worry for the woman and her family (Johnston et al., 1998). Additionally, the screening procedure, the mammogram, is not pleasant and many women find it uncomfortable and invasive, and some find it extremely painful (Sapir et al., 2003).

Screening for breast cancer does not change the incidence of cancer and no fewer women will develop breast cancer as a result of screening. Even without a screening programme, the same number of women will get cancer, although they may be diagnosed later by their GP after discovery of a lump or other symptoms. The NHS Information Centre (2012) produced a statistical report on the results of the breast screening programme in England using routinely collected data. These data are summarised in Table 3.2 for the age range 50-70 years. In terms of the cancers detected, other estimates have suggested that as many as half of these cases would have been picked up by other means regardless of screening participation (Independent UK Panel on Breast Cancer Screening, 2012).

### Table 3.2: Risks to women of breast cancer

<table>
<thead>
<tr>
<th>Outcome of screening</th>
<th>Risk per 1,000 women screened</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>≈108/1,000 screened</td>
</tr>
<tr>
<td>Benign biopsy</td>
<td>≈80/1,000 screened</td>
</tr>
<tr>
<td>Cancer detected</td>
<td>≈8/1,000 screened</td>
</tr>
<tr>
<td>Invasive cancer &lt;1.5cm detected</td>
<td>≈4/1,000 screened</td>
</tr>
</tbody>
</table>

*Source: The NHS Information Centre Screening and Immunisations (2012)*

Some of the possible outcomes of cancer screening are described in Figure 3.1, which shows the progression of women who choose to participate in screening and the relative benefits and harms. This diagram aimed to represent a sufficient simplification of reality and is therefore not exhaustive. Other benefits and risks are not incorporated in this diagram, nor are longer term outcomes which would overcomplicate the figure and lead to a Markov-style model of the probabilities of different states and potential routes (Taylor, 2010).

The benefits and risks of participating in a national breast screening programme has sparked a debate in the medical literature, which caused a complete review of the UK’s screening policy by the Independent UK Panel on Breast Cancer Screening in 2010, who
published their report in 2012. Whether a NHSBSP causes more harm than good continues to be debated by academics in leading medical journals (Baum, 2013; Kirwan, 2013). To contribute to the debate, there have been efforts to qualitatively discover what women consider when they weigh up the benefits and risks, research into better communication, and pushes for further involvement of women in decision-making (Hersch et al., 2013). By analysing a DCE of benefit-risk trade-offs in breast screening, and how these are affected by risk formats, this thesis adds to these important and on-going debates about the merits of a NHSBSP.
Figure 3.1: Progression of women through screening

- **Woman aged 50-70 invited for screening**
  - Mammography (x-ray of breasts)
    - Negative Result
      - NO CANCER
    - Positive Result
      - Yes: Not Cancerous
        - NO CANCER
      - Yes: Cancerous (fast-growing)
        - CANCER
      - Yes: Cancerous (slow-growing)
        - CANCER
  - Biopsy required?
  - Treatment offered?
  - Outcome
    - Woman finds out she does not have cancer
    - Woman has a biopsy but no cancer is found
    - Woman has treatment for a fast growing cancer and is saved by screening
    - Woman receives treatment for cancer so slow it would never have caused harm

Benefits of screening:
- **2 weeks after mammogram to results**

Risks of screening:
3.4.3 Benefit-risks trade-offs in breast screening
The decision to participate in screening is based on the individual’s perception of the advantages and disadvantages of screening and the risks of foregoing the mammogram. In the UK, women are invited for screening from the age of 50 years but as of 2011, there are still half a million women in the eligible age range who have been invited for screening but have not participated (The NHS Information Centre Screening and Immunisations, 2012); possibly because they do not believe the benefits outweigh the risks.

Screening women more frequently and for a larger age range, would increase the detection rate at the expense of increased risk of over-diagnosis and radiation-induced cancer (Tabár et al., 1987; Moss et al., 2006). Understanding women’s preferences for breast screening and how they balance the associated benefits and risks would help decision-makers provide an effective screening programme which maximises value.

3.4.4 Healthcare setting
The NHS currently screens for breast cancer in the community, with women invited by a letter sent their home address to attend their local screening centre, which may be a clinic, a mobile unit or a local hospital. Few women are referred to the breast screening service (<5% in 2010-2011) and most screening takes place in a primary care setting (The NHS Information Centre Screening and Immunisations, 2012).

3.4.5 Potential outputs of the DCE
As risk is an inherent component of any healthcare good or service, the decision to both approve the funding and use of services on the NHS involves service commissioners weighing up the risks and benefits. As discussed in section 3.4.2, there has been a well-publicised debate about whether the harms outweigh the benefits of breast screening. Quantification of women’s preferences through DCEs could contribute to this debate. The elicitation of women’s benefit-risk trade-offs in breast screening could provide a number of possible outputs including a measurement of their tolerance of risk in the programme and their WTP for the current screening programme.

Recent UK policy recommendations for breast screening have suggested expanding the screening age so it is available to both younger and older women. The DCE data could identify the drivers of demand and how subgroups of women could be targeted to improve uptake.
3.4.6 Existing DCEs relevant to cancer screening: a rapid review

The next section summarises the published DCE literature which has investigated preferences for cancer and screening generally, and breast screening programmes where available. The DCEs were identified from a systematic review conducted as part of this thesis (see Appendix 5.1 and section 5.2 of Chapter Five).

The results of the systematic review presented in Appendix 5.1 did not identify any studies that have elicited women’s preferences for the UK’s breast screening programme. However, studies of preferences for similar screening programmes have successfully used DCEs to elicit values for the associated benefit-risk trade-offs. This rapid review involved hand-screening the studies identified by the systematic review presented in Appendix 5.6 to find DCEs looking at preferences for breast cancer-related healthcare or diagnostic screening. For the purpose of this rapid review, Public Health England’s definition of screening was used:

“A process of identifying apparently healthy people who may be at increased risk of a disease or condition. They can then be offered information, further tests and appropriate treatment to reduce their risk and/or any complications arising from the disease or condition.”

(Public Health England 2013b)

The rapid review identified 17 relevant studies related to breast cancer or screening programmes. Two additional studies were identified (X1 and X11) which were not found in the systematic review. X1 was a study in Japanese so did not meet the English-language inclusion criterion, however, an English language abstract was provided. X11 is a working paper which was not peer-reviewed. The reviewed studies are summarised in Table 3.3. The complete list of references for the included studies is listed in Appendix 3.4. The review aimed to identify studies relevant to large-scale screening programmes in cancer rather than selective or multiphasic screening where the attributes were vastly different to breast screening and therefore uninformative for this case study.

Two DCEs were identified relating to breast screening (X1 and X11). X1 was in Japanese and could not be translated and X11 looked at clinicians’ preferences for referring women for screening outside of the current programme’s coverage. Three other studies were identified relating to breast cancer that considered women’s preferences for surgery (X2 and X4) and their preferences for genetic counselling (X3). The generic breast cancer studies were only tenuously linked to the chosen case study and had little transferable information relevant to the design of a DCE for this PhD.
In the identified screening DCEs, there were commonly occurring attributes such as frequency of screening, cost of the programme, test accuracy, discomfort or pain from screening, risk reduction from participation, and location of test provider. The most commonly occurring risk attributes referred to the programmes’ sensitivity or specificity, featuring in nine of the DCEs (X6, X7, X8, X10, X12, X14, X16, X17 and X18). In addition to specificity, the risk was communicated as: chance of a false-positive; chance of being recalled; chance of an unnecessary treatment; chance of an unnecessary follow-up; or risk of unnecessary colonoscopies. The sensitivity of the screening intervention was also described as chance of a false-negative; missed polyps; and possibility of missing a cancer. In line with the findings of Harrison et al. (2014), risk was most commonly communicated as a percentage.

The most common choice question for the screening DCEs was a ternary choice (used by ten studies: X5, X7, X8, X9, X10, X11, X13, X14, X15 and X19) with respondents choosing from two programmes or neither. The neither options were defined as either ‘No Screening’, indifferent or status-quo. The screening DCEs mainly elicited preferences from members of the public although three studies (X11, X12 and X14) administered their DCE to clinicians.

The review highlighted the potential originality of using breast screening in the UK as a case study for the DCE (only one other study (X11) had a UK setting). The review also identified some common designs in existing cancer screening-related DCEs such as the types of attributes, choice set designs and samples, which are useful starting points for consideration in the design of a breast screening DCE.
Table 3.3: Breast cancer and screening DCEs

<table>
<thead>
<tr>
<th>ID</th>
<th>Condition</th>
<th>Attributes</th>
<th>Risk format</th>
<th>Choice Set</th>
<th>Sample</th>
<th>Quantitative Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>DCE for mammography</td>
<td>Total amount of time taken for the screening Degree of breast pain Possibility of breast cancer being missed during the screening</td>
<td>Effectiveness of reducing deaths caused by breast cancer Total cost required for the screening</td>
<td>Not applicable</td>
<td>301 members of the public (Japan)</td>
<td>Conditional logit</td>
</tr>
<tr>
<td>X2</td>
<td>Follow-up breast cancer surgery</td>
<td>Attendance at an educational group programme Frequency of visits</td>
<td>Waiting time Contact mode Type of healthcare provider</td>
<td>Not applicable</td>
<td>331 breast cancer patients (Netherlands)</td>
<td>Random parameter logit</td>
</tr>
<tr>
<td>X3</td>
<td>Genetic counselling</td>
<td>Information Preparation</td>
<td>Surveillance Direction</td>
<td>Not applicable</td>
<td>210 female breast cancer study participants (Australia)</td>
<td>Random effects probit</td>
</tr>
<tr>
<td>X4</td>
<td>Breast reconstruction</td>
<td>Materials for reconstruction Number and duration of operations Short-term complications</td>
<td>Long-term complications Aesthetic result Waiting time</td>
<td>Percentages</td>
<td>386 breast cancer patients (Netherlands)</td>
<td>MNL</td>
</tr>
<tr>
<td>X5</td>
<td>Barrett esophagus surveillance</td>
<td>Number of times tested over 10 years</td>
<td>Reduction in risk of dying from esophageal carcinoma</td>
<td>Percentages</td>
<td>247 study patients (Netherlands)</td>
<td>Conditional logit</td>
</tr>
<tr>
<td>X6</td>
<td>Colorectal cancer screening</td>
<td>Test process Test frequency Requirement for a follow-up test if the initial screening test is positive Test-related pain or discomfort</td>
<td>Preparation for the test Risk of complications Test accuracy as measured by sensitivity Test accuracy as measured by specificity</td>
<td>Mix: Percentages for sensitivity and specificity, and frequencies (1/X) for risk of complications.</td>
<td>1588 members of the public (Canada and the United States of America (USA))</td>
<td>Bivariate probit</td>
</tr>
<tr>
<td>X7</td>
<td>Colorectal cancer screening</td>
<td>Deaths prevented Unnecessary colonoscopies</td>
<td>Result notification</td>
<td>Frequencies (Y/10,000 deaths prevented and X colonoscopies /death prevented)</td>
<td>301 members of the public (Australia)</td>
<td>Random effects probit</td>
</tr>
<tr>
<td>X8</td>
<td>Cervical cancer screening</td>
<td>Time between smears Waiting time for results Chance of being recalled</td>
<td>Chance of abnormality Chance of dying from cervical cancer Cost of the smear</td>
<td>Percentages</td>
<td>641 screened and unscreened mixed members of the public (Scotland).</td>
<td>Nested logit</td>
</tr>
<tr>
<td>ID</td>
<td>Condition</td>
<td>Attributes</td>
<td>Risk format</td>
<td>Choice Set</td>
<td>Sample</td>
<td>Quantitative Model</td>
</tr>
<tr>
<td>------</td>
<td>----------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>------------------------------</td>
<td>-------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>X9</td>
<td>Cervical cancer screening and vaccination</td>
<td>Lifetime risk of cervical cancer Lifetime risk of genital warts Frequency of smear testing Need for vaccine booster</td>
<td>Target group to vaccinate Risk of side effect Cost of the vaccine</td>
<td>Percentages</td>
<td>Ternary choice (including opt-out)</td>
<td>1157 member of the public (Canada)</td>
</tr>
<tr>
<td>X10</td>
<td>Prostate cancer screening</td>
<td>Risk reduction of prostate cancer-related death Screening interval Risk of unnecessary biopsies</td>
<td>Risk of unnecessary treatments Out-of-pocket costs</td>
<td>Simple graphics and frequencies (X/1000)</td>
<td>Ternary choice (including opt-out)</td>
<td>427 member of the public (Netherlands)</td>
</tr>
<tr>
<td>X11</td>
<td>Breast screening</td>
<td>Breast cancer risk Patient age Comorbidities</td>
<td>Physical functionality Cognitive functionality</td>
<td>Qualitative descriptor (normal/raised)</td>
<td>Ternary choice (screen/no screen/can’t decide)</td>
<td>139 clinicians (UK)</td>
</tr>
<tr>
<td>X12</td>
<td>Colorectal cancer screening</td>
<td>10-year mortality reduction after screening Risk of false negative Risk of false positive</td>
<td>Annual remuneration Number of avoided deaths-per 100,000 Information</td>
<td>Percentages and frequencies (X/10,000)</td>
<td>Binary choice with constant alternative</td>
<td>294 clinicians (France)</td>
</tr>
<tr>
<td>X13</td>
<td>Colorectal cancer screening</td>
<td>Reduction in mortality, Frequency of screening Complication risk Location</td>
<td>Duration of screening Patient preparation Pain from screening</td>
<td>Percentages and qualitative descriptors (‘small risk of complication’)</td>
<td>Ternary choice (including opt-out)</td>
<td>276 screened and unscreened mixed members of the public (Netherlands)</td>
</tr>
<tr>
<td>X14</td>
<td>Cervical cancer screening</td>
<td>Screening interval GP descriptors Time since previous screen Doctor’s recommendation</td>
<td>Incentive payment Chance of false negative Chance of false positive</td>
<td>Frequencies (1/X)</td>
<td>Ternary choice (2 different pap tests, standard and liquid, and an opt-out)</td>
<td>382 members of the public and clinicians mix (Australia)</td>
</tr>
<tr>
<td>X15</td>
<td>Colorectal cancer screening</td>
<td>Test type Screening interval</td>
<td>Mortality risk reduction</td>
<td>Percentages and icon arrays</td>
<td>Ternary choice (including opt-out)</td>
<td>1034 screened and unscreened mixed members of the public (Netherlands)</td>
</tr>
<tr>
<td>X16</td>
<td>Colorectal cancer screening</td>
<td>Test accuracy – find/miss cancers Test accuracy – find/miss polyps Unnecessary colonoscopies</td>
<td>Cost Preparation Collection</td>
<td>Frequencies (X/100)</td>
<td>Binary choice</td>
<td>1157 members of the public (Australia)</td>
</tr>
<tr>
<td>X17</td>
<td>Colorectal cancer screening</td>
<td>Test process Preparation Pain</td>
<td>Specificity Sensitivity Cost</td>
<td>Percentages</td>
<td>Binary choice (could state after if would opt-out)</td>
<td>547 members of the public (Canada)</td>
</tr>
<tr>
<td>ID</td>
<td>Condition</td>
<td>Attributes</td>
<td>Risk format</td>
<td>Choice Set</td>
<td>Sample</td>
<td>Quantitative Model</td>
</tr>
<tr>
<td>-----</td>
<td>----------------------------</td>
<td>-------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>X18</td>
<td>Colorectal cancer screening</td>
<td>Provider Sample type Sensitivity</td>
<td>Risk of unnecessary colonoscopy Test result provider Cost to you</td>
<td>Percentages and fractions</td>
<td>Binary choice with constant alternative</td>
<td>656 members of the public (France)</td>
</tr>
<tr>
<td>X19</td>
<td>Colorectal cancer screening</td>
<td>Preparation Location Pain</td>
<td>Risk of complication Frequency Duration</td>
<td>Percentages and icon arrays</td>
<td>Ternary choice (including opt-out)</td>
<td>280 screened and unscreened mixed members of the public (Netherlands)</td>
</tr>
</tbody>
</table>
3.5 Breast screening: designing the DCE

The next section of this chapter will describe the key stages of creating a DCE building on the step-wise approach following ISPOR guidelines (see Bridges et al. 2011b) for good research practice.

3.5.1 Research questions

This PhD sought to understand if, and how, the communication of risk attributes in a DCE affects respondents’ choices and the valuations derived. The case study for breast screening led to a specific set of defined research questions. The DCE aimed to elicit women’s preferences for breast screening and a quantification of their balance of the benefits and risks associated with screening programmes. The results of the DCE will contribute to the debate discussed in section 3.3 and the comparisons of different risk communication formats will investigate whether preferences are susceptible to change with the format used.

3.5.2 Sampling frame

There is extensive debate in the literature about whose preferences should be used in order to measure the values people place on different aspects of healthcare. The sample population who will answer the DCE survey must be appropriate to answer the research question and large enough to allow for subgroup analysis if necessary (Ryan & Watson 2011). A key criticism of published DCEs has centred around the characteristics of the respondents whose preferences were elicited (Bryan & Dolan 2004). It has been acknowledged that patients’, the public’s and health professionals’ preferences will differ and this should be carefully considered when designing a DCE (Montgomery & Fahey 2001). The sample chosen will also relate back to the perspective taken and whether an extra-welfarist or welfarist viewpoint is pursued.

In health economics more generally, it has been argued that the general public’s (tax-payers’) preferences should be taken into account for a publicly-funded health service (Bryan & Dolan 2004). Other contrasting views suggest that the patient experience plays an important role in determining preferences and that clinicians also have views to be incorporated (de Wit et al., 2000). DCEs have elicited preferences from healthcare professionals, hospital managers or policy makers (Bech, 2003; Gidman et al., 2007; Payne et al., 2011). There is some consensus (based on advice from the Panel on Cost-Effectiveness in Health and Medicine) that preferences taken from the public contain a broad range of views thus making them most informative for policy makers in healthcare (Weinstein et al., 1996).
Although breast cancer affects both men and women, the screening programme in the UK is only targeted at women. This is because of the considerably higher prevalence rate in women (Anderson et al., 2010). It is therefore, in this context, more appropriate to elicit preferences from women. A DCE eliciting preferences of men for a non-existent and unplanned screening programme would be neither useful, nor informative, for decision-makers.

In Table 3.3, one DCE (X5) was described which elicited preferences from people who were at an increased risk of developing cancer. However, this DCE was in the context of a service providing surveillance of a rare disease. As the NHSBSP routinely invites all women, it is difficult to rationalise why higher-risk women’s preferences would be more important.

Female members of the public are both tax-payers, and potential patients, because the NHSBSP as it currently stands will invite all women in England once they reach the age of 50 years. For the purpose of collecting preferences about breast screening, female members of the public were defined as the relevant study sample.

The key premise for the selected study sample is underpinned by guidelines published by NICE for the implementation of more commonly-used stated preference methods (time trade-off and EuroQol (EQ-5D)) that advocate preferences from a sample of the general public (NICE, 2013).

3.5.3 Identification of attributes and levels
The next stage in designing a DCE is to identify the characteristics of the good or service which drive demand. The identification of attributes and levels began as an iterative process of interviews with clinical experts (n=4), a patient representative (n=1) and female members of the public (n=4), and literature reviews based on the attributes identified in the rapid review of section 3.4.6.

3.5.3.1 Attributes
Breast screening is made up of positive and negative attributes which women must trade-off in their decision to attend the programme (see section 3.3). A DCE comprising few attributes was desirable so the effect of risk communication on choice could be investigated thoroughly while being cognisant of the potential cognitive burden of the task. A DCE needs to provide sufficient information about an intervention for respondents to make choices. The selection of attributes for the breast screening DCE balanced the requirement of a simple choice set with the need to generate informative preference data.
Commonly occurring attributes in existing cancer screening DCEs were identified (see section 3.4.6): test accuracy; discomfort or pain from screen; risk reduction from participation; location; test provider; cost; and frequency of the programme. Each of these identified attributes was discussed extensively with the supervisory team, breast screening experts and patients (n=5), and female members of the public (n=4).

The accuracy of the screening test was the most commonly occurring risk attribute included in the identified screening-related DCEs. Accuracy was frequently described in terms of the test’s specificity (risk of a false-negative) and sensitivity (risk of a false-positive). The specificity of a breast screening programme is the risk that the mammogram misses a true cancer. In the case of breast screening, the risk of missing a true cancer is vanishingly small and the problem of over-diagnosis is much greater. For this reason, test specificity was not included as an attribute in the DCE.

In breast screening, the sensitivity of the programme is most often described as a risk of either over-treatment or over-diagnosis. The definition of over-diagnosis varies as experts debate the inclusion of slow-growing cancers which are not always distinguishable from dangerous tumours and that of biopsies with negative results (Carter et al., 2015). In this DCE, the appropriate attribute selected was the risk of unnecessary treatment where women are called back for something which would not have harmed them and this could include repeat scans, biopsies which identify no cancer, and treatment for slow-growing tumours.

As mammograms are known to be uncomfortable (Sapir et al., 2003), pain would be an obvious attribute to include. However, pain is a difficult concept to communicate and suffers from subjective interpretation and although scales exist to communicate pain such as the Brief Pain Inventory (Cleeland & Ryan, 1992), it was felt that discomfort from screening would have induced another framing issue in addition to the risk attribute. In addition, the pain a woman might experience from having a mammography will be dependent on the equipment used, the skills of the radiographer and the woman’s physiology (Sapir et al., 2003). It was decided, therefore, that explaining what a mammogram involves, including the possible pain involved, in the training materials at the start of the DCE would be more beneficial than including pain as an attribute of a programme.

There is no potential reduction in the risk of developing breast cancer from participating in a breast screening programme. Screening women does not prevent the onset of breast cancer and the incidence rates of developing breast cancer are the same whether a women
is screened or not. Therefore an attribute for reduced risk of cancer is not appropriate. However, the long term health-related outcomes of breast cancer have been shown to change as a result of screening, with some studies concluding that women who participate in screening programmes are less likely to die of breast cancer, and they are also less likely to require a mastectomy (Berry et al., 2005).

The potential inclusion of an attribute to capture the risk of mortality from breast cancer was discussed with expert clinicians (n=4) at the Nightingale Centre, Wythenshawe Hospital, but they strongly believed it was not a relevant motivator of screening attendance. The experts stated that from their experience, women attended screening simply because they wanted to detect cancer but mortality was too long-term for them to seriously consider. For this reason, the probability of detecting cancer by screening was identified as an appropriate attribute to explain the benefits of breast screening. This attribute allowed investigation of a positive risk; the probability of a favourable event (a benefit) occurring.

Three of the screening DCEs identified in rapid review of section 3.4.6 included the process attributes relating to the provider of the programme or location of the programme (studies X14, X18 and X19). As the research questions relating to this DCE were to investigate the communication of risk attributes, the outcomes of screening were seen to be more important and therefore attributes relating to the provider were not included. It was feared that additional attributes may add to the cognitive load of respondents and potentially detract from the main attributes of interest (risk).

A cost attribute is useful to include in a choice experiment as it allows monetary valuations and welfare calculations. To understand if, and how, elicited valuations differ between risk communication formats; a cost attribute was included. Cost is a notoriously difficult attribute to frame when respondents are used to consuming healthcare free at the point of use (Johnson et al., 2011). The cost attribute was described as out-of-pocket expense of attending the screening programme. The cost attribute was framed with the aim of reducing the incidence of protest bids or zero-response bias.

Frequency of screens occurred as an attribute in nine of the screening DCEs identified in the rapid review of 3.4.6. The frequency of screens is potentially important to women as the procedure is uncomfortable and inconvenient. Preferences for screening intervals are also of interest to policy makers with trials examining the effect of biennial screening. DCE responses could help understand how this might affect demand (Moss et al., 2006). Including screening frequency was problematic in this DCE as it would be directly
correlated with the cost of screening attribute. Pre-pilot exploratory interviews with four female colleagues revealed that women were confused by the inclusion of screening frequency. The pilot respondents would try to identify the ‘best value’ programme by dividing or multiplying the cost by the frequency.

To reduce strategic behaviour, the programme was described as lasting twenty years (in line with current policy) with screening occurring five times over that period. This screening frequency allowed the cost to be presented over a lifetime (starting from the age of 50 years), and the cost per screen. Using this approach prevented women from making frequency-based calculations in the stated preference exercise.

The three attributes chosen for inclusion in the pilot study were:

1) Probability of detecting a cancer by screening over a lifetime
2) Risk of unnecessary treatment over a lifetime
3) Out-of-pocket cost of the programme (per screen and over a lifetime)

3.5.3.2 Levels

Assigning levels to these attributes describing the risks and benefits of screening, was challenging. The national breast screening programme has a high uptake rate which covers a large proportion of the population (The NHS Information Centre Screening and Immunisations, 2012). It is therefore difficult to calculate the baseline rates of benefits and risks for an unscreened population given the challenges of accounting for selection bias from women who opt-out. For example, the women who choose not to attend may be at particularly low risk or attend for mammograms privately. The range of levels used reflected estimates identified from the literature. The exact figures used in the levels depended on the perceived definition, population sample and whether the data were collected from randomised controlled trials or observational studies.

The attribute ‘probability of detecting a cancer’ was assigned levels based on a study by Welch & Frankel (2011). The actual value depends on age, race and family history. It was possible, however, to identify estimates for the average woman entering screening at aged 50 years with a 20 year probability of detecting a cancer set at about 3.5%. This estimate was calculated as the product of the risk of developing breast cancer (approximately 6%) and proportion of breast cancers found by mammography (approximately 65%).

The levels chosen for the attribute reflecting the probability of detecting a cancer by screening were: 3%, 7%, 10% and 14%. If women who perceived themselves to be at a
low risk of developing breast cancer, opted out of screening, then the proportion of breast cancers found by mammography could be underestimated and the 3.5% probability calculated by Welch & Frankel (2011) would be a conservative estimate hence representing the lower bound of the level range. A higher range for the level was set by reflecting detection rates achievable through stratified or more frequent screening (Hall et al., 2014).

The attribute ‘risk of unnecessary treatment’ was also assigned levels from a literature review. The levels selected represent the range of views that centre on the definition of what constitutes an over-treatment or over-diagnosis (see section 3.4.2). The Independent Review of Breast Cancer Screening (2012) estimated that “just over 1% of women invited for screening over a twenty year period would have an over-diagnosed cancer” (p.1778). This estimate only incorporated over-treatment for cancer and did not include women who underwent unnecessary biopsies. A review of the Norwegian screening programme estimated false recalls after mammography to be approximately 20% (Hofvind et al., 2004). The final levels used for the attribute reflecting the risk of unnecessary treatment were: 1%; 5%; 10%; and 20%.

Four levels of out-of-pocket costs were used: £20 per screen (£100 over a lifetime); £50 per screen (£250 over a lifetime); £150 per screen (£750 over a lifetime); and £200 per screen (£1,000 over a lifetime). The lower bound represents the lowest amount of out-of-pocket expenses, associated with taking time off work and travelling to a screening centre. The upper-bound was set as a realistic maximum based on the price of private mammograms in the UK (BreastHealthUK, 2014). The full list of attributes and levels selected for piloting in the DCE are shown in Table 3.4.

Table 3.4: Attributes and levels used in pilot DCE

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of detecting a cancer</td>
<td>The chance of detecting a cancer from screening over a 20 year period</td>
<td>3%; 7%; 10%; 14%</td>
</tr>
<tr>
<td>Risk of unnecessary treatment</td>
<td>The probability of being recalled for a procedure or procedures when no harm existed over a 20 year period</td>
<td>1%; 5%; 10%; 20%</td>
</tr>
<tr>
<td>Out-of-pocket cost of screening over a lifetime</td>
<td>The costs of attending the programme including original screens and recalls. These could include transport, time-off work, carer costs.</td>
<td>£100; £250; £750; £1,000</td>
</tr>
</tbody>
</table>
3.5.4 Choice format
There are advantages and disadvantages of labelling the alternatives in a DCE compared with using a generic (un-labelled) question format. Kruijshaar et al. (2009) established labelling alternatives in a DCE added realism to the choices but de Bekker-Grob et al. (2010) found that labelling a DCE detracted respondents’ attention from the attributes. In this thesis, the key research questions relate to the communication of risk attributes and anything that detracted from these would have been a detriment to the study. In addition, for this example, screening is conducted via a mammogram and there are no immediately apparent labels for different screening programmes. Therefore an unlabelled DCE was considered appropriate for this case study, and given the research question, thus the programmes took a generic (un-labelled) name of programme A or programme B.

Recent estimates of breast screening in the UK suggest uptake to the national programme is approximately 75% (The NHS Information Centre Screening and Immunisations, 2012). Whilst this is relatively high compared with other screening programmes such as bowel cancer (Weller et al., 2006), there is still a large number of women who are invited but do not attend. These uptake figures imply that some women have revealed ‘no demand’ for the programme, and consequently, a forced choice experiment would be inappropriate. Therefore an opt-out alternative of ‘No Screening’ was included in addition to the two programmes, A and B, in the pilot study.

3.5.5 Experimental design
The experimental design refers to the generation of alternatives and choices sets from a selection of attributes and levels. A full factorial design for a DCE of three attributes, each containing four levels would result in 64 possible profiles. This would be too many alternatives for any one person to consider, and would likely result in respondent fatigue. Therefore, the DCE aimed to use a fractional factorial design but still be capable of creating precise parameter estimates.

Using all the possible combinations of levels can create many profiles, and too many profiles can be expensive to distribute, demanding for respondents and time-consuming to analyse (Ryan & Watson 2011). To reduce this full factorial to a more manageable size, the profiles can be blocked or fractional factorials can be designed (Street et al., 2005). Blocking involves splitting the profiles into different chunks and creating multiple questionnaires (Ryan et al., 2008). However, if one questionnaire block suffers from a particularly poor response rate, it can lead to an inconclusive analysis (Viney et al., 2002; Witt et al., 2009).
Reducing a full factorial design to a fractional factorial design may involve assuming that all interactions are negligible or zero, and therefore only the main effects are estimated. It could also allow for a select few interactions to be present rather than estimating them all (Viney et al., 2005). A fractional factorial could be created by selecting profiles at random from the full factorial, but more systematic methods that ensure efficiency are usually employed to ensure an orthogonal or D-efficient design (de Bekker-Grob et al., 2012).

In orthogonal design plans, there is level balance as all attribute levels occur with equal frequency and there is no correlation between the attributes (Street & Burgess, 2007). This level balance allows identification of the effect of variables in levels of all effects independently as the attributes of the design are statistically independent, and minimises the variance in the parameters (Ryan & Watson 2011). These inherent characteristics are present in the fractional factorial design plans available in the Hahn & Shapiro (1966) catalogue.

If there are unequal numbers of levels for each attribute or the combination does not appear in the Hahn & Shapiro (1966) catalogue, then it is possible to overestimate and ‘collapse’ at the expense of level balance (Ryan & Watson 2011). If three attributes have four levels and one attribute has two levels, then it is possible to follow the design for a four attribute, four level plan with the two level attribute appearing twice as often, at the expense of level balance.

The other option is to employ a D-efficient design. D-efficiency involves minimising the inverse of the variance-covariance matrix in a maximum likelihood estimation (Johnson et al., 2013), and often involves the use of a software algorithm (using programmes such as Ngene®) to search for a design which maximises this efficiency (de Bekker-Grob et al., 2012). Previously, researchers have generated D-efficient designs assuming that parameters are equal to zero, however, using a Bayesian approach and incorporating prior assumptions (expected values) to improve statistical efficiency is a recent advance that is increasingly being used (de Bekker-Grob et al., 2012).

Ultimately, D-optimal designs (based on minimising the D-error) are a balance between statistical and response efficiency (Johnson et al., 2013). However, when designing a DCE there is a trade-off in allowing for estimation of effects thought to be important, and a design efficient in terms of allowing estimates from a smaller sample size. The method chosen will depend on the research question and the resources available for the study.

From the profiles suggested by the mathematical design, the choice sets of alternatives for the DCE need to be created. A recent systematic review by Clark et al. (2014) found most
authors now use software to design their DCEs. One popular package, Ngene®, is a mathematical software used to design choice experiments for given design criteria (such as orthogonality, level balance, blocking or validity tests). The software is then able to calculate the most ‘efficient’ design depending on the criteria set (for given sample sizes and model specification). Ngene® identifies the design from algorithms that use nested Monte Carlo simulation to select the most efficient design after running an extensive number (~2 million) of iterations where each iteration produces a design with the pre-defined number of choice situations. The software automatically compares each design produced from each iteration using the D-error, computed on the basis of expected values (priors) of the predicted model parameters.

The pilot DCE was constructed using a Bayesian efficient design. This involved defining priors for each of the main and interaction effects predicted to be in the final estimation. As neither the main effects nor interactions were perfectly known, point estimates were specified with a normal distribution in the design algorithm explained in the software syntax (see Figure A1 Appendix 3.5). The software package Ngene® (Choice Metrics, 2012) was used to generate a design minimising the D-error for a MNL. The initial pilot design in Ngene® used 2,000 Halton (random) draws to sample from the distribution of priors. This approach incorporated the uncertainty around the priors, and allowed deviation from the expected values whilst still ensuring efficiency of the design was maximised.

For the pilot study, four blocks of ten choices were chosen as ten choices seemed an appropriate starting amount for the number of questions to use in the pilot study. de Bekker-Grob et al. (2012) found over three-quarters of DCEs used fewer than 16 choice sets to reduce the cognitive burden of the exercise. Ngene® randomly selected from the full factorial design (64) which was defined by the number of attributes (three) and levels (four). The pilot design can be found in Table A2, Appendix 3.5.

3.5.5.1 Validity tests

The reliability of DCE data may be demonstrated by proving responses are consistent and respondents have made informed choices (Telser & Zweifel, 2007). However, validity of data can be tested through tests within the choice set and by comparisons with revealed preference data in internal and external validity checks, respectively.

One way of ensuring the results of a DCE are internally valid is through the inclusion of tests within the survey. Internal validity tests involve the development of specific choice sets to check that the respondents’ choices exhibit the properties of: transitivity where if A is better than B, and B is better than C, then A is preferred to C; monotonicity where if A is
preferred to B and B is at least as good as C, then A is also preferred to C; stability where if A is preferred to B in one set, then A is preferred to B in another set too; and dominance when the respondent chooses an option that is superior in all ways (Johnson et al., 2009; Varian, 1992).

Tests for behaviour in line with these axioms can be incorporated in the design of a DCE to check for its internal validity. One of the fundamental and most common tests is for monotonicity (de Bekker-Grob et al., 2012). Each of the blocks in the DCE contained a check for internal validity to verify whether the respondent was answering in line with expectations and theory underpinning monotonicity. The test included a choice-set where one alternative had a higher probability of detection a cancer; a lower risk of unnecessary treatment; and a lower cost of screening. Failure of the test would not necessarily mean the respondent was irrational, but could instead indicate further explanation about the health intervention or choice experiment was required.

3.5.6 Survey materials
The DCE choice questions were completed as part of an online survey which included information (training materials) about the screening programme and an explanation of the forthcoming task. After the choice task, the respondents were asked for background information which might be useful to understand if, and how, different women gave different responses. The development of the preceding training materials is described in section 3.5.6.1 and the background questions are explained in section 3.5.6.2.

3.5.6.1 Training materials: pilot study
The information preceding the choice questions in a healthcare DCE survey is an important stage of informing the respondents about the disease area and proposed good or service. Whilst the information should be accessible to all potential respondents, it should also be interesting enough to keep their attention. As a starting point, breast screening pamphlets distributed as part of the NHSBSP were examined to extract appropriate terminology and phrasing (Slaytor & Ward, 1998; Cancer Research UK, 2013). Publically-available information resources on Public Health England’s website (Public Health England, 2013a) were also consulted to create a comprehensive yet simple introduction to breast screening and the choice task. The information was supplemented with Figure 3.1 to show women the potential progression through screening and the related risks and benefits in a visual way.
3.5.6.2 Background questions: pilot study

Factors which could influence women’s preference for breast screening or their interpretation of risk, were collected for possible use in a subgroup analysis or as covariates. Socio-demographic questions were taken from a review of established surveys (such as the Census and Understanding Society) which informed the context, question frame and answer categories (see Appendix 3.6).

In addition to socio-demographic questions, the survey also assessed respondents’ numeracy by using the standardised questions identified by the structured review described in section 3.2 and Appendix 3.1 (see Figure A3). In addition to questions on numeracy, their experience of probabilities and their perception of risk in general were captured through Likert scale questions. Examples of the additional questions can be found in the pilot DCE shown in Appendix 3.8. The order of the questions in the survey reflects their importance. As respondents could drop out, after an initial screen to confirm the participants wished to continue, and a verification of their age for sampling, the first questions respondents answered were the choice sets.

The background questions were then ordered by their perceived importance for the analysis to capture as much data as possible if a respondent dropped out. Socio-demographic questions came first and questions on their feelings about probability came last.

3.5.7 Ethical approval

On the 18th September 2013, ethical approval was granted by The University of Manchester Research Ethics Committee (see Appendix 3.7).

3.5.8 The piloting process

After receiving ethical approval, a preliminary survey was created using online Sawtooth software (Sawtooth, 2012). The software provides an alternative to website development by providing a platform to enable survey webpages and choice set-style questions. The survey was hosted online using The University of Manchester’s secure servers and databases to store survey responses. The software’s inbuilt functions for design and analysis were not used in this PhD due to the availability of specialised, and more sophisticated, packages such as Ngene®. Starting with the pilot DCE (see Appendix 3.8), an iterative piloting process was used. The stages of piloting described in sections 3.5.8.1 to 3.5.8.2 were concurrent processes of changing and re-testing. After approximately ten drafts and re-drafts, a satisfactory final DCE was developed.
3.5.8.1 Experts and healthcare professionals

The first stage of piloting involved confirming that the attributes and levels selected were appropriate for a breast screening DCE. Whilst the design stages involved expert consultation, the complete pilot was also tested with experts at the Nightingale Centre, Wythenshawe Hospital and genetic counsellors at St Mary’s Hospital, Manchester Royal Infirmary (n=4). In addition, a link to an online version of the survey was distributed to attendees (n=11) of a Genesis Breast Cancer meeting held in December 2013 in Manchester.

The discussion with four experts identified a need for a change in terminology from ‘unnecessary treatment’ to ‘unnecessary follow-up’, as they thought this was a more accurate description of the recall process. The experts also suggested including a set of validated questions on risk perception and concerns about developing breast cancer (see Evans et al. 1994; Evans et al. 1993; Hopwood et al. 2001). The importance of religious questions and feelings of fate, already incorporated in the questionnaire, were emphasised as important by the experts.

3.5.8.2 Qualitative interviews

In-depth face-to-face interviews using the think-aloud method (see Chapter Five) were conducted with female members of the public (n=4) and a patient representative (n=1). Women were recruited with an online advertisement placed on The University of Manchester’s Research Ethics Website (see Appendix 3.9). Women who expressed an interest in the study were given an information sheet (see Appendix 3.11) to read over before the interview was arranged. As a thank you, the interviewees were given a £10 Amazon voucher after the interview was completed. The interviews took place on The University’s Main Campus once the women consented (see Appendix 3.12 for the consent form). Contact with the patient representative was made via a clinical expert involved in the piloting.

The aims of these interviews were two-fold: 1) to pilot the DCE survey (see Appendix 3.8); and 2) to test the think-aloud interview schedule (see Appendix 3.10). Although interviews were recorded, the qualitative data collected was not formally transcribed and instead the field notes were used to make changes to the DCE and the preliminary interview schedule.

It became immediately apparent that the younger members of the public were not familiar with the screening programme and the procedure of a mammogram. Therefore a more
interactive resource was sought to maintain the respondents’ attention. The NHS Choices’ website video describing the screening programme was selected to add into the training material for the main DCE survey. The video was edited to remove any persuasive language which may have swayed women towards screening, and instead provided a factual account of the process. Use of the video was approved by NHS Choices (see Appendix 3.13).

The ‘No Screening’ option appeared to be ignored by the interviewees in piloting. A change to the DCE was made. The selected levels for the ‘No Screening’ option were explicitly stated in the main DCE survey: a zero probability of detecting a cancer through screening; a zero probability of having an unnecessary follow-up from screening; and no cost, if the respondent opted out.

It was also apparent some interviewees found thinking-aloud difficult. After contacting an author of a think-aloud study identified in the systematic review presented in Appendix 5.1 (personal communication with Dr Sudeh Cheraghi-Sohi in July 2013), a warm-up exercise was included in the main think-aloud study. The warm-up exercise asked respondents how many windows they had in their house, which helped them to verbalise their thoughts, even if it felt unnatural.

The pilot interviews also provided an insight into the practicalities of recruitment and to an approximate interview length, which was around 45 minutes.

3.5.8.3 Internet panel: pilot study

To rapidly acquire responses to test the priors used in the experimental design, the DCE was piloted with an internet panel. The panel involved an initial sample of fifty-six women who were recruited via the internet panel provider, ResearchNow®. The sample was restricted to women aged 50 years and over to acquire information about the perspectives of older women not captured by the pilot interviews. The pilot internet panel also allowed for potential technical complications associated with hosting an online survey and collecting data to be identified and eliminated prior to the main survey. For example, the generation of different pass-in/pass-out IDs; survey redirect links for participant payments; and an idea of the ResearchNow® referral and response rate.

The pilot data were analysed using conditional logit and heteroscedastic conditional logit models. Details of the results of the internet panel pilot can be found in Appendix 3.14. All coefficients were significant and had the expected signs. However, a feature of the design
meant linearity of preferences in risk and detection could not be tested as a 0% level only appeared in the ‘No Screening’ option. This was changed in the main survey design.

The alternative specific constant (ASC) on ‘No Screening’, represent the mean of the distribution of the unobserved effects in the random component, $\epsilon$, was insignificant in the pilot study. Free-text comments from the survey indicated that some respondents were giving protest responses that occurred in women who objected to the idea of paying for screening. Other women were sceptical of the purpose of the survey, possibly because they were used to completing surveys for private companies. Therefore, the main survey reiterated that screening would remain free on the NHS and that the survey was being used for the purpose of a PhD research study. A few women also commented that they were unsure what the word ‘attribute’ was referring to. The numeracy questions suffered from a particularly low response rate with many women skipping these questions (n=16, 26%).

A 0% level was introduced into the attribute risk of unnecessary follow-up. It was not introduced into the detection rate as this would generate inferior alternatives with no benefit from screening and yet potential costs (both financial and to health). The updated priors from the initial sample were then used to generate a new experimental design (with the additional level) following the steps described in section 3.5.5 which only required two blocks of ten choice sets. The final design included eleven choice sets including an additional dominant choice as a test of internal validity.

The new design was tested in a second pilot study using a sample of 58 women collected via internet panel provider ResearchNow® and also included defined levels for the ‘No Screening’ option. The training materials were changed to include the NHS video described in section 3.5.8.2, and it was explained explicitly that answers would be used only for university research. An additional question about women who had experienced screening and the cost incurred to them was added to provide information about real-life out-of-pocket costs incurred. The use of the word ‘attribute’ was removed and replaced with ‘characteristic’, and a survey logic was added which required women to give a reason for skipping the numeracy questions. The second internet panel experimental design, survey and results can be found in Appendix 3.15, 3.16 and 3.17, respectively. The second internet panel survey resulted in no additional changes to the survey or design.

**3.5.9 DCE design: main study**

The final attributes, attribute definitions, levels and specification of the opt-out alternative are shown in Table 3.5.
Table 3.5: Attributes and levels used in final DCE

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition</th>
<th>Levels for programmes</th>
<th>Levels for opt-out of ‘No Screening’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of detecting a cancer</td>
<td>The chance of detecting a cancer from screening over a 20-year period</td>
<td>3%; 7%; 10%; 14%</td>
<td>None: no cancers detected (0%)</td>
</tr>
<tr>
<td>Risk of unnecessary follow-up</td>
<td>The probability of being recalled for a procedure or procedures when no harm existed over a 20-year period</td>
<td>0%; 1%; 5%; 10%; 20%</td>
<td>None: no unnecessary follow-ups (0%)</td>
</tr>
<tr>
<td>Out-of-pocket cost of screening over a lifetime</td>
<td>The costs of attending the programme including original screens and recalls. These could include transport, time-off work, carer costs.</td>
<td>£100 (£20 per screen); £250 (£50 per screen); £750 (£150 per screen); £1,000 (£200 per screen)</td>
<td>None: no cost to you (£0)</td>
</tr>
</tbody>
</table>

An example of the final choice sets in each of the survey versions can be found in Figure 3.2, where risk was communicated as either a percentage or as a percentage and icon array.

Appendix 3.15 and 3.16 shows the experimental design and the DCE survey subsequently used in the empirical studies described in Chapters Four, Five and Six.
Figure 3.2: Example choice questions

If these were your only options, which would you choose?
Choose by clicking one of the buttons below:
(1 of 11 choice questions)

<table>
<thead>
<tr>
<th></th>
<th>Programme A</th>
<th>Programme B</th>
<th>No Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers detected by screening</td>
<td>3%</td>
<td>14%</td>
<td>None: no cancers detected</td>
</tr>
<tr>
<td>Women who will have an unnecessary follow-up</td>
<td>1%</td>
<td>0%</td>
<td>None: no unnecessary follow-up</td>
</tr>
<tr>
<td>Out-of-pocket cost to you of screening programme per screen</td>
<td>£20 per screen (£100 over your lifetime)</td>
<td>£50 per screen (£250 over your lifetime)</td>
<td>None: no cost to you</td>
</tr>
</tbody>
</table>

If these were your only options, which would you choose?
Choose by clicking one of the buttons below:
(1 of 11 choice questions)

<table>
<thead>
<tr>
<th></th>
<th>Programme A</th>
<th>Programme B</th>
<th>No Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers detected by screening</td>
<td>3%</td>
<td>14%</td>
<td>None: no cancers detected</td>
</tr>
<tr>
<td>Women who will have an unnecessary follow-up</td>
<td>1%</td>
<td>0%</td>
<td>None: no unnecessary follow-ups</td>
</tr>
<tr>
<td>Out-of-pocket cost to you of screening programme per screen</td>
<td>£20 per screen (£100 over your lifetime)</td>
<td>£50 per screen (£250 over your lifetime)</td>
<td>None: no cost to you</td>
</tr>
</tbody>
</table>
3.6 Conclusion

Eliciting women’s preferences for the benefits and risks of a NHSBSP was identified to be a useful case study for this thesis. Breast screening was found to be topical and well-debated, suggesting results from the DCE could be useful for decision-makers seeking to make changes to a breast screening programme. The topical nature also meant that it was a familiar disease area, which diminished the challenge of communicating an obtuse clinical topic to DCE respondents.

The DCE was extensively piloted with experts and female members of the public. The final pilot indicated that the attributes and levels selected were appropriate and the priors acquired informed an efficient design for the final study. Pilot work also ensured that the DCE was accessible to the public, confirming they understood the subject matter and the choice task. In the pilot, all attributes were estimated to be significant in women’s decision to participate in screening and also had coefficient signs in line with expectations. The DCE design allowed for investigation of how a positively-framed risk (benefit) compared with a negatively-framed risk. The inclusion of the cost attribute allowed for valuation of these risk attributes using marginal WTP (MWTP). The following three chapters (Chapters Four, Five and Six) describe the empirical studies designed to investigate the effect of risk communication on individual’s preferences and decision-making heuristics.
Chapter Four
Benefit-risk trade-offs for breast screening: results of the DCE

4.1 Introduction
This chapter uses econometric regression methods to quantify women’s preferences for the benefit-risk trade-offs associated with breast screening and assess the effect of risk communication format on these preferences. The chapter addresses three main objectives: 1) to identify the trade-offs women are prepared to make in a NHSBSP using the DCE designed in Chapter Three; 2) to investigate whether the framing of risk affects their benefit-risk trade-offs made; and 3) identify the presence and determinants of preference heterogeneity. The chapter starts with a description of the key research questions in section 4.2. In section 4.3, the methods for: 1) data collection; and 2) the analysis are described. In section 4.4, the results of the DCE study are presented, and in section 4.5 the key findings, implications and study limitations are detailed. The chapter concludes in section 4.4.

4.2 Aims and objectives
The aim of the large scale DCE study was to identify if, and how, the communication of risk attributes affect women’s preferences for breast screening. The sample used the final DCE designed in Chapter Three which can be found in Appendix 3.15. The design included 11 choice sets, with a test for internal validity, and a two-block design generated using Ngene® (Choice Metrics, 2012). To investigate whether the risk communication format had an effect, the design included two experimental conditions: percentages only (PO); or icon arrays and percentages (IAP). The study sought to address a number of key research questions:

1) What is the structure of women’s utility for breast screening?
2) Is the randomness of choice affected by the framing of the risk attributes?
3) After accounting for randomness, do individual’s preferences differ with the framing of risk?
4) Do groups of women have similar preferences in the form of latent-classes?
5) What are the predictors of class membership?
6) What are the marginal rates of substitution (MRS) between the attributes?
7) How do changes in risk affect the probability of a woman attending screening?

The research questions were answered using a variety of methods discussed in section 4.3.
4.3 Methods
The following sections describe the methods used to elicit and analyse preference data. The recruitment of the study sample is described in section 4.3.1. The models chosen to answer the research questions stated in section 4.2 are described in section 4.3.2. In section 4.3.3, other analyses are presented including calculations to determine predicted uptake and the MRS between attributes.

4.3.1 Data collection
An online survey was designed using Sawtooth software version 8.3.8 (Sawtooth, 2012) and uploaded to the internet as described in Chapter Three. Women were recruited to complete the survey via an internet panel provider, ResearchNow®. Internet panels were identified as a suitable source of achieving high quality yet low cost responses, with studies showing that this approach yields reliable response data efficiently (Ščasný & Alberini, 2012; Mulhern et al., 2015). Women who completed the survey received ‘points’ from the internet panel provider which they could accumulate and exchange for gift vouchers.

The first page of the survey (as shown in Appendix 3.16) explained the task and provided the opportunity for the respondents to opt-out of the study. The women were then randomly allocated to receive the DCE with either the PO or IAP risk format. After the desired sample of women was acquired, the survey was closed to prevent new entrants.

4.3.1.1 Recruitment
As a guide to DCE sample size, Orme (1998) suggested the power calculation presented in equation 4.1. The required sample size depends on the number of choice tasks (T), the number of alternatives in a choice set (A), and largest number of levels in any attribute (l):

\[ \text{Sample size} > \frac{500l}{TA} \quad [4.1] \]

In this DCE, the Orme (1998) power calculation would suggest a minimum sample size of 152 (allowing for tests of the two risk communication formats). Despite the calculation by Orme (1998) suggesting a sample size of 152, this study aimed for a much larger sample. The large sample size was in response to literature suggesting preferences for healthcare are highly heterogeneous; even more so than for other areas of DCE research (such as transport and environment) (Fiebig et al., 2010). As a result, the sample size was constrained only by financial resources and instead the aim was to achieve as many views and choices as possible.
Female members of the public were identified as an appropriate sample in section 3.5.2 of Chapter Three. Quotas for each of the age bands selected for the survey (see p.351 of Appendix 3.16) were used to select a sample. These age bands were identified after the discussions with experts (n=4) described in section 3.5.3 of Chapter Three. The views of women currently in screening (who may have better formed preferences) and women about to enter screening (with regards to upcoming changes in policy) were felt to be of particular interest to the screening experts interviewed. Therefore the last two age bands (45 to 49 years and over 50 years) were over-sampled.

4.3.1.2 Data storage and transformation

The choice data generated were anonymous and stored as the survey progressed using The University of Manchester’s secured server. After the sample was collected, the data were downloaded and transformed from comma-separated values (CSV) files into STATA® (StataCorp, 2011) .dta files and LatentGold® (Statistical Innovations, 2013) .lgf files.

4.3.1.3 Coding

The data were coded to rescale some attributes, such as cost, to make the coefficients more easily interpretable by reducing the number of decimal places. In addition, both effects coding (EC) and dummy coding (DC) were used, depending on the model specification.

Dummy coding involves dropping a level of an attribute and transforming the other levels into individual (dummy) variables which are equal to one when the level is present or zero if it is absent. The effects of the dummies are then comparable to the dropped level, although the effect of this level is consequently captured by the constant term and so the estimated coefficients are consequently correlated with it (Bech & Gyrd-Hansen, 2005). Effects coding allows for more information to be gained about the dropped level and eliminates the identification problem that occurs when the dropped level is incorporated within the intercept (Bech & Gyrd-Hansen, 2005). Effects coding allows the attribute interaction effects to be independently estimated whereas dummy variables would result in artificially high interactions and main effects because the utility associated with the dropped level is not distinguishable from the other elements of utility also incorporated in the intercept estimate (Ryan et al. 2008; Bech & Gyrd-Hansen 2005). The coding of the attribute levels are shown in Table 4.1.
Table 4.1: Attributes, levels and coding

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Label</th>
<th>Levels</th>
<th>Coding</th>
<th>EC1</th>
<th>EC2</th>
<th>EC3</th>
<th>EC4</th>
<th>DC1</th>
<th>DC2</th>
<th>DC3</th>
<th>DC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of detecting a cancer</td>
<td>Detect</td>
<td>3%</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7%</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10%</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14%</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Risk of unnecessary follow-up</td>
<td>Risk</td>
<td>0%</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1%</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5%</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10%</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Out-of-pocket cost of screening</td>
<td>Cost</td>
<td>£100</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>£250</td>
<td>2.5</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>£750</td>
<td>7.5</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>£1000</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

EC: effects coding
DC: dummy coding

Other dummy variables were generated for key items of interest and to simplify some complex response options. For example, ‘To what extent are you concerned about your own risk of breast cancer?’ gave women five options (not at all, a little, quite a lot, very much, no idea) so a dummy variable was generated to identify women who stated that they were not at all concerned. A dummy was also created to identify respondents who had failed the internal validity test, and an additional dummy variable was generated to identify the risk format the respondent had received. The dummy variables generated and their definitions and coding and presented in Table 4.2.

Table 4.2: Dummy variable definitions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAP</td>
<td>Identifier for risk format respondent received</td>
<td>1 = IAP&lt;br&gt;0 = PO</td>
</tr>
<tr>
<td>Wrong</td>
<td>Identifier for respondents who failed the internal validity test by choosing the inferior alternative.</td>
<td>1 = Failed test&lt;br&gt;0 = Passed test</td>
</tr>
<tr>
<td>Unconcerned</td>
<td>Dummy variable collapsed from question: ‘To what extent are you concerned about your own risk of breast cancer?’ to identify respondents who stated ‘Not at all’ about their risk of cancer.</td>
<td>1 = Not concerned about risk&lt;br&gt;0 = Concerned/No idea</td>
</tr>
<tr>
<td>Employed</td>
<td>Dummy variable collapsed from question: ‘What is your occupational status?’ to identify respondents in work (either employed full-time, part-time or self-employed).</td>
<td>1 = Employed&lt;br&gt;0 = Not employed</td>
</tr>
<tr>
<td>White</td>
<td>Dummy variable collapsed from question: ‘What is your ethnic group?’ to show White respondents (either ‘White British/Irish’ or ‘White Other’).</td>
<td>1 = White&lt;br&gt;0 = Non-White</td>
</tr>
</tbody>
</table>

4.3.2 Models

Descriptive statistics were used as a preliminary exploration of the DCE choice data. Respondent characteristics were initially explored to determine whether any differences had occurred, by chance or error, in the randomisation process to the different risk formats
or the two design blocks. After ensuring randomisation was successful, investigation of the research questions presented in section 4.2 began. The research questions were answered using multiple discrete choice models which are now explained.

The choice data were modelled using a random utility maximisation framework (McFadden, 1974). Using equation 4.2 as a starting point, an individual’s \((n)\) indirect utility \((U)\) for an alternative \((j)\) can be written as:

\[
U_{nj} = V_{nj} + \varepsilon_{nj} \quad [4.2]
\]

In equation 4.2, utility comprises a deterministic component \((V_{nj})\) and an unobservable, random component incorporating unobservable attributes \((\varepsilon_{nj})\). The probability of an individual \(n\) choosing alternative \(i\) over alternative \(j\) is given by equation 4.3:

\[
P_{ni} = \text{Prob}(U_{ni} > U_{nj} \text{ for all } j \text{ where } i \neq j) \quad [4.3]
\]

Assuming that \(\varepsilon_{nj}\) is iid extreme value with a variance \(\sigma^2\) then a closed-form choice probability can be given by equation 4.4:

\[
P_{ni} = \frac{\exp V_{ni}}{\sum_j \exp V_{nj}} \quad [4.4]
\]

Where \(V_{nj}\), the deterministic component of utility is explained by:

\[
V_{nj} = \beta X_{nj} \quad [4.5]
\]

In equation 4.5, \(\beta\) is a vector of parameters to be estimated and is the relative level of satisfaction for individual \(n\), and \(X_{nj}\) is a vector of observed attributes for alternative \(j\).

### 4.3.2.1 Discrete choice models

The first research question sought to understand the structure of women’s preferences for breast screening. This involved: 1) assessing whether the effects of attributes on utility had the expected sign; 2) investigating whether an ASC was needed for any alternative; and 3) investigating whether there existed non-linearity in women’s preferences for the risks associated with breast screening. To do so, a simple conditional logit model was used.

#### 4.3.2.1.1 Conditional logit model

Model 1 was a conditional logit model. The conditional logit model is a simple discrete choice model which makes a number of assumptions. Combining equations 4.4 and 4.5 means the conditional logit model can be given by:
The conditional logit model presented in equation 4.6 was used to conduct a preliminary investigation into the structure of women’s utility. It was hypothesised that the attributes ‘risk of unnecessary follow-up’ and ‘cost’ would have negative coefficients as more risk and more cost are intuitively seen as less desirable, and the attribute ‘probability of detection’ would have a positive coefficient. These a priori expectations did not influence the analysis, but provided a first step of checking that the DCE was theoretically valid through a simple model (Model 1A).

Conditional logit models were also used to make simple specification changes to the model. Equation 4.7 simply investigated the effect of each attribute on utility where Risk$_{nj}$, Detect$_{nj}$ and Cost$_{nj}$ were the exogenous attributes associated with alternative $j$, and $\varepsilon_{nj}$ the unobserved random component which has a variance and a set of covariances.

$$U_{nj} = \beta_1 \text{Detect}_{nj} + \beta_2 \text{Risk}_{nj} + \beta_3 \text{Cost}_{nj} + \varepsilon_{nj}$$ \hspace{1cm} \text{[4.7]}

In this DCE, it could be expected that there may be a difference in $\varepsilon_{nj}$ between the ‘No Screening’ option and Programmes A and B, as individuals may have gained utility from just participating in screening, beyond what was explained by the attributes. In equation 4.8, $\alpha_{n0}$, an ASC was introduced. This ASC represents the mean of the distribution of the unobserved effects in the random component, $\varepsilon_{nj}$, associated with the opt-out for individual $n$. This ASC can therefore be interpreted as the average effect of the unobserved screening attributes on the utility of not participating in a programme relative to any form of screening. Equation 4.8 was estimated using a conditional logit (Model 1B)

$$U_{nj} = \alpha_{n0} + \beta_1 \text{Detect}_{nj} + \beta_2 \text{Risk}_{nj} + \beta_3 \text{Cost}_{nj} + \varepsilon_{nj}$$ \hspace{1cm} \text{[4.8]}

In equation 4.9, ASCs were introduced for Programme A and Programme B to investigate any left-right bias in respondents’ choices.

$$U_{nj} = \alpha_{nA} + \alpha_{nB} + \beta_1 \text{Detect}_{nj} + \beta_2 \text{Risk}_{nj} + \beta_3 \text{Cost}_{nj} + \varepsilon_{nj}$$ \hspace{1cm} \text{[4.9]}

Equation 4.9 was also estimated using a conditional logit (Model 1C). A left (right) bias was present if $\alpha_{nA}$ was significantly higher (lower) than $\alpha_{nB}$.

The functional form of women’s risk preferences was investigated using Model 1D (a conditional logit estimation of equation 4.10). Assuming a linear utility function when preferences were non-linear may result in biased estimates (Torres et al., 2011). Therefore

$$P_{ni} = \frac{\exp(\beta x_{ni})}{\sum_j \exp(\beta x_{nj})} \hspace{1cm} \text{[4.6]}$$
a dummy for each level of risk was incorporated in the model. As a result, the utility from alternative \( j \) for individual \( n \) is:

\[
U_{nj} = \alpha_{n0} + \beta_1 Detect_{nj} + \beta_2 Risk\_1_{nj} + \beta_3 Risk\_5_{nj} + \beta_4 Risk\_10_{nj} + \beta_5 Risk\_20_{nj} + \beta_6 Cost_{nj} + \epsilon_{nj} \quad [4.10]
\]

Comparing Model 1D with Model 1B allowed an investigation into the linearity of women’s risk preferences. To do so, a likelihood ratio (LR) test was conducted, which tested whether Model 1B or Model 1D was best-fitting. If Model 1D was best-fitting, this would indicate that women’s risk preferences are non-linear. In addition, non-linearity was tested by plotting the coefficient of the dummy variables representing each level of risk in Model 1D. Investigation into non-linearities in preferences for the other attributes could not be conducted as the 0 levels (0% of detection and £0) only occurred in the ‘No Screening’ option and were therefore perfectly collinear.

4.3.2.2 Testing icon array effects

Research questions two and three were concerned with the effect of risk format on the randomness of respondents’ choices and on utility parameters (individuals’ preferences), respectively. Under the conditional logit model described in section 4.3.2.1, the error variance across individuals was assumed constant. That is to say, choice consistency was the same across choices (and the scale parameter was normalised to one). However, this DCE may have unequal error variance as respondents received different risk communication formats, depending on the DCE they were randomised to. This would mean that differences in estimated parameters in the conditional logit models across risk formats would not reflect true differences in preferences, as these differences would also capture differences in the scale parameter (choice consistency) across groups. To allow for this, investigations into scale heterogeneity were made in order to answer research question two: is the randomness of choice affected by the formatting of the risk attributes?

4.3.2.2.1 Impact of risk format on choices: preferences and scale.

The choice sets the respondents completed were identical in all attributes and experimental design resulting in no other context-specific components to the utility functions described in section 4.2.1.1. However, how the risk communication formats could have affected respondents’ decision rules. The structured review of risk communication formats presented in Appendix 3.10 found icon arrays as a preferable risk communication format which could translate into different stated preferences (as respondents’ can make choices
which better reflect their preferences as they are more informed) or differences in scale (as task complexity is reduced and, as a consequence, choice consistency improves).

To test if the risk communication format affected the utility parameters (preferences), the null hypothesis that preferences were the same between the two communication formats was $H_0: \beta_{PO} = \beta_{IAP}$, where $\beta_{PO}$ represents the estimate utility parameters for those receiving the percentages only risk format, and $\beta_{IAP}$ represents the utility parameters for those receiving icon arrays and percentages. However, estimation of the conditional logit models described in section 4.3.2.1.1 actually involved estimates of $\lambda_{PO}\beta_{PO}, \lambda_{IAP}\beta_{IAP}$ where $\lambda$ was a scale parameter confounded with the $\beta$ estimates. Therefore, with a standard conditional logit, it was not possible to simply compare model coefficients from the different risk formats as differences in parameter estimates could be due to: 1) differences in utility parameters (preferences) (the $\beta$s); 2) differences in consistency or scale factors (the $\lambda$s); or 3) both differences in utility and scale. Crucially, if the scale parameters differed between IAP and PO then the point estimates may not reflect real differences in preferences.

Plotting $\beta_{PO}$ against $\beta_{IAP}$ results in a gradient $\frac{\lambda_{IAP}}{\lambda_{PO}}$ (the ratio of the error variance for PO respondents to the error variance for IAP respondents). If the points lie on a straight line, this would indicate a ratio of one and suggest that the scale parameters are approximately equal i.e $\lambda_{PO} \approx \lambda_{IAP}$. However, the DCE used in this study had only three attributes (detect, risk, cost) which made parameter plots limited in their ability to confirm or disconfirm choice consistency (as there are only three data points to determine linearity). Therefore to understand whether choice consistency was affected by the risk communication format, heteroscedastic conditional logit models were used to estimate the size, direction and significance of the scale parameter. Introducing the scale parameter directly into the estimated utility function allowed for individual-level heteroscedasticity (variance in response error) as shown by equation 4.11:

$$P_{ni} = \frac{\exp(\lambda_n\beta X_{ni})}{\sum_j \exp(\lambda_n\beta X_{nj})} \quad [4.11]$$

In equation 4.12, $\lambda_n$ is a scale term which is allowed to vary by $IAP_n$, a dummy variable which equals unity if individual $n$ received the IAP version of the DCE:

$$\lambda_n = \exp(\delta IAP_n) \quad [4.12]$$

$\delta$ reflects the influence of the risk communication format on the error variance, and therefore the scale parameter. As shown in equation 4.12, for identification, the baseline
scale parameter is set equal to one for respondents completing the PO version of the DCE. Testing if $\delta = 0$ is a test of constant error variance across respondents from all versions. Accounting for scale heterogeneity in the heteroscedastic conditional logit means that any differences in estimated utility parameters across risk formats reflects true differences in preferences.

Therefore, to understand how the risk communication format affected any of the marginal utilities associated with an attribute, the data was pooled and equations 4.13 and 4.12 were estimated using a heteroscedastic conditional logit model (Model 2A):

$$U_{nj} = \alpha_{no} + \lambda_n \beta_1 Detect_{nj} + \lambda_n \beta_2 Risk_{nj} + \lambda_n \beta_3 Cost_{nj} + \lambda_n \beta_4 Detect_{nj} IAP_n + \lambda_n \beta_5 Risk_{nj} IAP_n + \epsilon_{nj} \quad [4.13]$$

In equation 4.13, $Detect_{nj} IAP_n$ and $Risk_{nj} IAP_n$ represent interactions terms between the probability of detecting a cancer attribute and the probability of unnecessary follow-up, with the dummy variable indicating the risk format received, respectively. No such interaction term was included for the cost attribute, as cost was framed in an identical manner in both versions of the DCE. As a result, to investigate differences in preferences due to the risk format (and therefore to answer research question three), the following hypothesis was tested:

$$H_0: \beta_4 = \beta_5 = 0 \quad \text{and} \quad H_1: \beta_4, \beta_5 \neq 0.$$

To investigate whether risk format affected choice consistency (and therefore to answer research question two), the following hypothesis was tested:

$$H_0: \delta = 0 \quad \text{and} \quad H_1: \delta \neq 0.$$

To investigate whether respondents who failed the internal validity test had different preferences and/or choice consistency, equations 4.14 and 4.15 were estimated using a heteroscedastic conditional logit model (Model 2B):

$$U_{nj} = \alpha_{no} + \lambda_n \beta_1 Detect_{nj} + \lambda_n \beta_2 Risk_{nj} + \lambda_n \beta_3 Cost_{nj} + \lambda_n \beta_4 Detect_{nj} IAP_n + \lambda_n \beta_5 Risk_{nj} IAP_n + \lambda_n \beta_6 Detect_{nj} Fail_n + \lambda_n \beta_7 Risk_{nj} Fail_n + \lambda_n \beta_8 Cost_{nj} Fail_n + \epsilon_{nj} \quad [4.14]$$

$$\lambda_n = \exp(\delta IAP_n + \varphi Fail_n) \quad [4.15]$$

Equation 4.14 is identical to equation 4.13 but with additional interaction terms between all attributes and a dummy for whether an individual failed the internal validity test, $Fail_n$, and equation 4.15 is identical to equation 4.12, but with $Fail_n$ included as an additional predictor of scale. Hypotheses $H_0: \beta_6 = 0$, $H_1: \beta_6 \neq 0$; $H_0: \beta_7 = 0$, $H_1: \beta_7 \neq 0$; and $H_0: \beta_8 = 0$, $H_1: \beta_8 \neq 0$, were then tested to test whether failing the internal validity test affected women’s preferences for the detection, risk and cost attributes, respectively, and
\( H_0: \phi = 0 \) and \( H_1: \phi \neq 0 \) tested to test whether those who failed had less consistent choices.

All estimations of Models 1 and 2 were conducted in STATA® using maximum likelihood estimation (MLE) to acquire parameter estimates.

### 4.3.2.3 Investigating further preference heterogeneity through preference groups

Research questions four and five were concerned with establishing whether groups of women have similar preferences, and whether observables describing women different preference groups. The conditional and heteroscedastic conditional logit models presented previously were aggregated models which may hide underlying variation in preferences. The estimation of latent-class models allowed identification of subgroups with similar preferences by investigating preference heterogeneity amongst responses to the DCE.

The conditional and heteroscedastic conditional logit models are also based on the independent of irrelevant alternatives (IIA) assumption. Under IIA, it is assumed that the ratio of choice probabilities for any two alternatives is unaffected by the presence or absence of another alternative (Louviere et al., 2000). For example, assume an individual is faced between choosing a doctor and a hospital nurse, and that this respondent chooses between these options with equal probability, 0.5. If another option, say district nurse, was introduced the IIA assumption requires that this equal probability must be maintained, so the probability of choosing a doctor, district nurse and hospital nurse must all be 0.33. However, if the introduction of this new alternative resulted in probabilities of 0.5, 0.25 and 0.25 respectively, the IIA assumption would be invalid. In addition to allowing the investigation of further preference heterogeneity, latent-class models also relax this IIA assumption.

#### 4.3.2.3.1 Latent-class analysis

The third estimated model (Model 3) was a latent-class model. In latent-class models, respondents are allowed to be probabilistically assigned to classes which differ in their utility functions. The choice probabilities and the marginal utilities are dependent on the class, \( c \). The latent-class model is therefore specified as:

\[
P_{n|i|c} = \frac{\exp(\beta_c X_{ni})}{\sum_{j=1}^J \exp(\beta_c X_{nj})} \quad [4.16]
\]

In equation 4.16, \( P_{n|i|c} \) is the probability that an individual, \( n \), chooses alternative \( i \), given that this individual is in class, \( c \). Class membership is modelled in MNL form:
\[ P_{nc} = \frac{\exp(Z_n \Phi)}{\Sigma_{c=1}^{C} \exp(Z_n \Phi)} \]  \hspace{1cm} [4.17]

In equation 4.17, \( Z_n \) is a vector of individual respondent characteristics included as explanatory variables. In this study, details about the respondent characteristics were collected in the final questions of the survey (see Appendix 3.16, p. 368-385). The questions were chosen because they were believed to potentially influence a respondent’s choices and were therefore incorporated as preference-class covariates. Estimates \( \Phi \) indicate which of these covariates characterise difference preference groups. In addition, the \( IAP_n \) dummy, which represents the risk communication format received by the respondent, was also included as a preference-class covariate. Identification was achieved by imposing \( \Sigma \Phi = 0 \) so:

\[ P_n = \Sigma_{c=1}^{C} P_{nc} P_{n|c} \]  \hspace{1cm} [4.18]

Equation 4.18 was estimated using MLE. The number of classes must be imposed exogenously, thus the total number of preference-classes, \( C \), was varied and each model compared using information criteria to select the best fitting model via a calibration process. Information criteria are measures of the goodness of fit for maximum likelihood estimation models. The most widely used information criteria are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and Consistent Akaike Information Criterion (CAIC). These criteria vary in their penalties (such as for sample size, number of parameters) so different information criteria can support different models depending on their emphasis for sensitivity (having enough parameters to model relationships) or specificity (over-fitting) (Dziak et al., 2012). The results of the different information criteria can all be interpreted in the same way: the model with the lowest comparative value, the better (absolute values are meaningless). There is debate around which information criterion is preferable (Dziak et al., 2012) and models were also compared using the predictive error (how many choices were correctly predicted by the model).

4.3.2.3.2 Scale-adjusted latent-class analysis

The latent-class model was then extended to allow for each latent-class to comprise subsets of respondents with the same preferences but different error variances (choice consistency).

As in the conditional logit model, if the scale parameter differed between risk communication formats, then a standard latent-class model may not reveal true preference heterogeneity, as differences in estimated parameter across groups will reflect both
differences in preferences and differences in the scale parameter (choice consistency). Scale-adjusted latent-class models were used to overcome this. Scale-adjusted latent-class models capture differences in the scale parameter by explicitly grouping respondents with similar choice consistencies. The scale-adjusted latent-class model was used to identify latent preference-classes (Pc) within which different scale-classes (Sc) with different scale parameters, \( \lambda \), were present. Each of these Sc are associated with a scale-membership probability. Incorporating the scale parameter back into equation 4.16, yields equation 4.19:

\[
P_{ni|c,s} = \frac{\sum_{s=1}^{S} \pi_s \exp(\lambda_s \beta_c X_{ni})}{\sum_{j=1}^{J} \exp(\lambda_s \beta_c X_{nj})}
\]  \[4.19\]

Equation 4.19 was estimated using a scale-adjusted latent-class model (Model 4). \( \pi_s \) is the scale membership probability with \( 0 \leq \pi_s \leq 1, \sum_{s=1}^{S} \pi_s = 1 \). Variables which could predict choice consistency could be included to predict scale-class membership. In this study, self-reported task difficulty and risk format were included as Sc covariates. The scale-adjusted latent-classes were estimated using the software LatentGold® (Statistical Innovations, 2013) and the number of scale and preference-classes were selected through comparisons of the BIC, AIC and CAIC. These information criteria were also used to compare these models against standard latent-class models.

### 4.3.3 Marginal rates of substitution

Research question six was concerned with MRS between attributes. The results of Models 1 to 4 provided information which could be used to investigate how respondents were prepared to trade between attributes through their MRS. The parameter estimates can be interpreted in terms of willingness to exchange money (WTP) or WTA risk. Using the cost attribute, the MWTP for particular changes in non-cost attributes was estimated by dividing the coefficient of the attribute of interest by the coefficient of the cost attribute. The MWTP for the attributes probability of detecting a cancer and risk of unnecessary follow-up were calculated using equations 4.20 and 4.21, respectively.

\[
MWTP \ for \ a \ 1\% \ point \ increase \ in \ detection = \frac{\beta_{\text{Detect}}}{-\beta_{\text{Cost}}}
\]  \[4.20\]

\[
MWTP \ for \ a \ 1\% \ point \ decrease \ in \ risk \ of \ unnecessary \ follow-up = \frac{-\beta_{\text{Risk}}}{-\beta_{\text{Cost}}}
\]  \[4.21\]

The generated WTP values were then used to estimate an overall value of the screening programme. As identified in Chapter Three, there are challenges in identifying the current levels of each attribute in the NHSBSP as it currently stands due to extensive debates (see
section 3.4.2 of Chapter Three). Therefore, predicted WTP for a screening programme was calculated for different scenarios as detailed in Table 4.3.

Table 4.3: WTP for hypothetical screening programmes

<table>
<thead>
<tr>
<th>Probability of detecting a cancer</th>
<th>Screening worst-case</th>
<th>Screening intermediate</th>
<th>Screening best-case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of detecting a cancer</td>
<td>3%</td>
<td>10%</td>
<td>14%</td>
</tr>
</tbody>
</table>

The MRS between the attributes ‘probability of detecting a cancer’ and ‘risk of unnecessary follow-up’ show the benefit-risk trade-offs women were prepared to make for an NHSBSP. The ratio of coefficients of the probability of detection and risk of unnecessary follow-up \( \frac{\beta_{\text{Detect}}}{\beta_{\text{Risk}}} \) represents the number of unnecessary follow-ups women are willing to accept for an additional cancer detected, per 100 women screened.

Research question seven sought to understand the external validity of the DCE. As there was no revealed preference data, uptake was estimated for the different screening scenarios presented in Table 4.3 and compared to current levels of screening uptake.

4.4 Results
This next section presents the results of the large scale internet panel DCE and utilises the Models 1 to 4 to answer the research questions stated in section 4.2.

4.4.1 Descriptive statistics
The final study comprised 1,000 women who completed the DCE survey entirely. 1,018 women completed the DCE choice sets, but 18 women did not finish all the background questions. Out of the 1,018 women who completed the choice questions, 507 people received the PO version and 511 people received the IAP version. Out of the 1,000 women who completed the DCE entirely, 498 people received the PO version and 502 people received IAP version.

It was difficult to attribute an exact response rate to an internet panel survey. However, 38,251 emails were sent from ResearchNow® and 23,082 of these were delivered. 2,205 women clicked on the link to the survey. 511 women clicked on this when their age quota was full and another 60 clicked on the link after the survey had closed. The response rate in terms of survey loads as a proportion of emails delivered was therefore around 9%. A full breakdown of the exact response rate can be found in Appendix 4.1.
Table 4.4 shows the sample characteristics for a number of key variables, for all respondents and the IAP and PO subsamples. A more detailed breakdown of the sample characteristics can be found in Appendix 4.2. No significant differences occurred in the characteristics between the two risk communication formats. A logistic regression was also used to check if any of the covariates predicted risk format membership but none were statistically significant (see Appendix 4.3). The computerised randomisation to each survey version of the DCE survey appeared to be successful.

Table 4.4: Sample characteristics

<table>
<thead>
<tr>
<th>Age group</th>
<th>Overall</th>
<th></th>
<th></th>
<th></th>
<th>PO</th>
<th></th>
<th></th>
<th></th>
<th>IAP</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>18-24</td>
<td>104</td>
<td>10.2</td>
<td>44</td>
<td>8.7</td>
<td>60</td>
<td>11.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-34</td>
<td>103</td>
<td>10.1</td>
<td>51</td>
<td>10.0</td>
<td>52</td>
<td>10.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35-44</td>
<td>253</td>
<td>24.9</td>
<td>127</td>
<td>25.0</td>
<td>126</td>
<td>24.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45-49</td>
<td>253</td>
<td>24.9</td>
<td>131</td>
<td>25.8</td>
<td>122</td>
<td>23.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td>305</td>
<td>30.0</td>
<td>154</td>
<td>30.4</td>
<td>151</td>
<td>29.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Occupational Status**

| Employed full-time | 343 | 34.0% | 167 | 33.3% | 176 | 34.7% |
| Employed part-time | 211 | 20.9% | 114 | 22.7% | 97 | 19.1% |
| Self-employed     | 58  | 5.8%  | 31  | 6.2%  | 27  | 5.3%  |
| Unemployed/ Long-term sick/Laid off | 68 | 6.7% | 30 | 6.0% | 38 | 7.5% |
| Other: e.g Retired/ House wife | 329 | 32.6% | 160 | 31.9% | 169 | 33.3% |

**Education**

| Less than 4 O-levels | 150 | 14.9% | 74 | 14.7% | 76 | 15.0% |
| 5+ O-levels          | 112 | 11.1% | 54 | 10.8% | 58 | 11.4% |
| NVQs/ A-levels       | 308 | 30.5% | 165 | 32.9% | 143 | 28.2% |
| Undergraduate degree | 270 | 26.8% | 129 | 25.7% | 141 | 27.8% |
| Master’s/ PhD        | 99  | 9.8%  | 52 | 10.4% | 47 | 9.3%  |
| Other qualification  | 70  | 6.9%  | 28 | 5.6%  | 42 | 8.3%  |

**4.4.2 Internal validity**

As described in section 3.5.5.1 of Chapter Three, an internal validity test was included with a dominant choice option. In this study, 9% of the total sample failed this test. As shown in Table 4.5, out of the 90 respondents who answered incorrectly, 42 received the PO version and 48 received the IAP; this difference was not statistically significant (p=0.533). In
addition to the validity test, there was no statistically significant difference between the two formats with regards to self-reported task difficulty (p=0.640).

Table 4.5: Validity test failure and self-reported task difficulty.

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th></th>
<th>PO</th>
<th></th>
<th>IAP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Failed</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>8.8%</td>
<td>42</td>
<td>8.3%</td>
<td>48</td>
<td>9.4%</td>
</tr>
<tr>
<td>Task difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very easy</td>
<td>346</td>
<td>34.2%</td>
<td>172</td>
<td>34.2%</td>
<td>174</td>
<td>34.2%</td>
</tr>
<tr>
<td>Easy</td>
<td>299</td>
<td>29.6%</td>
<td>155</td>
<td>30.8%</td>
<td>144</td>
<td>28.3%</td>
</tr>
<tr>
<td>Neither easy nor difficult</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult</td>
<td>97</td>
<td>9.6%</td>
<td>45</td>
<td>9.0%</td>
<td>52</td>
<td>10.2%</td>
</tr>
<tr>
<td>Very difficult</td>
<td>10</td>
<td>1.0%</td>
<td>6</td>
<td>1.2%</td>
<td>4</td>
<td>0.8%</td>
</tr>
</tbody>
</table>

With confidence that the randomisation procedure was successful, discrete choice modelling of the choice data continued, starting with the conditional logit models.

4.4.3 Investigating the functional form of utility
The conditional logit model results are shown in Table 4.6. The results of Model 1A show the attribute coefficients from the conditional logit were in line with \textit{a priori} expectations with cost and risk having negative and statistically significant coefficients, whereas detection was positive and statistically significant. The negative coefficient on risk indicates that the higher the level of this attribute in an alternative, the less likely an individual was to choose this alternative. The positive coefficient for detection indicates that the women preferred a programme with higher levels of this attribute.

Table 4.6: Results of the conditional logit Models 1A, 1B and 1C

<table>
<thead>
<tr>
<th></th>
<th>Model 1A</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model 1C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASC(on A)</td>
<td>1.588***</td>
<td>(0.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASC(on B)</td>
<td>1.534***</td>
<td>(0.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASC(on none)</td>
<td>-1.569***</td>
<td>(0.05)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detect</td>
<td>0.141***</td>
<td>(0.00)</td>
<td>0.080***</td>
<td>(0.00)</td>
<td>0.079***</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.022***</td>
<td>(0.00)</td>
<td>-0.047***</td>
<td>(0.00)</td>
<td>-0.046***</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.036***</td>
<td>(0.00)</td>
<td>-0.108***</td>
<td>(0.00)</td>
<td>-0.107***</td>
<td>(0.00)</td>
</tr>
<tr>
<td>N</td>
<td>1018</td>
<td></td>
<td>33594</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p<0.005; ** p<0.01; ***p<0.001; standard errors in parentheses
In Model 1B, the ASC for the option ‘No Screening’ was found to be statistically significant and negative. This suggests respondents were more likely to choose either A or B over no screening at all, indicating that women gained utility from attending screening beyond that explained by the attribute levels. In Model 1C the ASCs on Programme A and B indicate respondents were likely to opt-in but were no more likely to choose A over B, or *vice versa*, and indicates no left-right bias was present. A likelihood ratio test showed Model 1B offered better fit (p<0.01) than Model 1A, and thus further models incorporated the ASC for ‘No Screening’.

In Model 1D the continuous variable of risk was recoded into levels (see Table 4.1). The results of Model 1D presented in Table 4.7 showed the coefficient on risk increased by approximately equal intervals with overlapping confidence intervals implying preferences for risk were monotonically increasing and could consequently be considered linear. A likelihood ratio test showed that allowing for non-linearity in the risk attribute in Model 1D did not increase fit (p>0.05), and thus further models assumed that a linear representation of the attributes was compatible with the data.

**Table 4.7: Results of the conditional logit Model 1D**

<table>
<thead>
<tr>
<th></th>
<th>Model 1D</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC(on none)</td>
<td>-1.624*** (0.06)</td>
</tr>
<tr>
<td>Detect</td>
<td>0.074*** (0.00)</td>
</tr>
<tr>
<td>Risk</td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>-0.030 (0.04)</td>
</tr>
<tr>
<td>5%</td>
<td>-0.289*** (0.06)</td>
</tr>
<tr>
<td>10%</td>
<td>-0.355*** (0.05)</td>
</tr>
<tr>
<td>20%</td>
<td>-0.962*** (0.05)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.111*** (0.00)</td>
</tr>
<tr>
<td>N</td>
<td>1018</td>
</tr>
<tr>
<td>Obs</td>
<td>33594</td>
</tr>
</tbody>
</table>

*p<0.005; **p<0.01; ***p<0.001; standard errors in parentheses*

In addition to the LR test, the parameter estimates were plotted which visually confirmed an approximately linear relationship (see Figures 4.1 to 4.3). A linear relationship was assumed for subsequent investigations.
Figure 4.1: Functional form of the risk attribute for DCE respondents

Figure 4.2: Functional form of the detection attribute for DCE respondents

Figure 4.3: Functional form of the cost attribute for DCE respondents
With confidence that the functional form was appropriately specified, more sophisticated models were estimated to investigate the presence of scale and preference heterogeneity to answer subsequent research questions.

### 4.4.4 Testing risk format effects

Estimation results from the heteroscedastic conditional logit models (Model 2A and 2B) are presented in Table 4.8. In Model 2A, all attribute coefficients were statistically significant and had the expected signs, again consistent with *a priori* expectations. All coefficients relating to the interactions between attributes and risk format, and the coefficient of risk format in the scale term were insignificant, suggesting that the risk format did not lead to any direct differences in either marginal utility or choice consistency, and that the format of risk had no effect in this study sample.

**Table 4.8: Heteroscedastic conditional logit results**

<table>
<thead>
<tr>
<th></th>
<th>Model 2A</th>
<th>Model 2B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Standard Error</td>
</tr>
<tr>
<td>ASC(on none)</td>
<td>-1.497***</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Detect</td>
<td>0.081***</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.047***</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.103***</td>
<td>(0.01)</td>
</tr>
<tr>
<td>IAP*detect</td>
<td>-0.010</td>
<td>(0.01)</td>
</tr>
<tr>
<td>IAP*risk</td>
<td>0.005</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Fail*detect</td>
<td>-0.086***</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Fail*risk</td>
<td>0.094***</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Fail*cost</td>
<td>0.222***</td>
<td>(0.02)</td>
</tr>
<tr>
<td>IAP</td>
<td>0.094</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Fail</td>
<td>0.063</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Fail*cost</td>
<td>-0.029</td>
<td>(0.13)</td>
</tr>
<tr>
<td>N</td>
<td>1018</td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>33594</td>
<td></td>
</tr>
</tbody>
</table>

* * p<0.005; ** p<0.01; ***p<0.001; standard errors in parentheses

However, in Model 2B the coefficients relating to interactions between attributes and failure of the validity test suggested respondents who failed the internal validity test made a significant difference to the parameter estimates. A likelihood ratio test comparing Models 2A and 2B showed Model 2B had statistically significant (p<0.001) better fit. However, responses of those who failed were included in the subsequent stages of analysis in line with recommendations from the literature (see Hougaard et al. 2012; Lancsar & Louviere 2006) and because their responses were no more random (with a statistically insignificant effect of ‘Fail’ on the scale parameter).

### 4.4.5 Investigating preference heterogeneity

With confidence that the data from each risk communication format could be pooled because of the insignificant scale and interaction terms for IAP in Model 2, latent-class models were estimated to explore preference heterogeneity further and answer research
question four and five. In Table 4.9, a summary of latent-class models with different numbers of preference-classes and their associated information criteria and log likelihood (LL) values are presented. On the BIC and CAIC, the 6-Pc model was preferred, however, the 7-class model was preferred based on the fact it minimised the AIC. The AIC is more sensitive (lower false-negative rate) but BIC is more specific (lower false-positive rate). In other studies, BIC has performed better than AIC (which tended to over-fit the data) in tests (Nylund, 2007). Therefore, on the basis of the BIC result, a 6-Pc model provided the best fit.

Table 4.9: Information criteria for varying preference-class numbers in Model 3

<table>
<thead>
<tr>
<th>Number of preference classes</th>
<th>LL</th>
<th>BIC(LL)</th>
<th>AIC(LL)</th>
<th>CAIC(LL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Pc</td>
<td>-9925.1295</td>
<td>19877.922</td>
<td>19858.2591</td>
<td>19881.922</td>
</tr>
<tr>
<td>2-Pc</td>
<td>-7954.8521</td>
<td>16186.3331</td>
<td>15989.7042</td>
<td>16226.3331</td>
</tr>
<tr>
<td>3-Pc</td>
<td>-7485.4462</td>
<td>15496.4875</td>
<td>15122.8925</td>
<td>15572.4875</td>
</tr>
<tr>
<td>4-Pc</td>
<td>-7042.2923</td>
<td>14859.1456</td>
<td>14308.5846</td>
<td>14971.1456</td>
</tr>
<tr>
<td>5-Pc</td>
<td>-6804.6897</td>
<td>14632.9065</td>
<td>13905.3795</td>
<td>14780.9065</td>
</tr>
<tr>
<td><strong>6-Pc</strong>*</td>
<td>-6661.3985</td>
<td><strong>14595.29</strong></td>
<td>13690.7969</td>
<td><strong>14779.29</strong></td>
</tr>
<tr>
<td>7-Pc</td>
<td>-<strong>6541.0412</strong></td>
<td>14603.5415</td>
<td><strong>13522.0823</strong></td>
<td>14823.5415</td>
</tr>
</tbody>
</table>

*Chosen model

In the 1-Pc model, the prediction error was 0.4227, which meant that just over half (57.7%) of choices were predicted correctly. In the 6-Pc model, the prediction error reduced to 0.1906 with over 80% of choices accurately predicted. Details of 6-Pc Model 3 can be found in Appendix 4.4.

4.4.6 Accounting for scale heterogeneity

The latent-class models presented in section 4.4.5 did not account for differences in scale, which could produce biased class parameter estimates or even suggest an incorrect number of latent-classes, C. Fitting a scale-adjusted latent-class model, allowing for groups of respondents to have different error variances, allowed for further analysis of preference heterogeneity without the confounding scale factor.

A second calibration procedure to identify the correct number of Sc and Pc was conducted. In Table 4.10, the different class models and associated information criteria are presented. When the latent-class analysis was run with scale-classes, the 3-Sc-6-Pc model had both the lowest BIC and lowest CAIC.
Table 4.10: Information criteria for varying class numbers in latent-class and scale-adjusted latent-class models.

<table>
<thead>
<tr>
<th>Number of classes</th>
<th>LL</th>
<th>BIC(LL)</th>
<th>AIC(LL)</th>
<th>CAIC(LL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Pc Model</td>
<td>-9925.13</td>
<td>19877.92</td>
<td>19858.26</td>
<td>19881.92</td>
</tr>
<tr>
<td>2-Pc Model</td>
<td>-7954.85</td>
<td>16186.33</td>
<td>15989.7</td>
<td>16226.33</td>
</tr>
<tr>
<td>3-Pc Model</td>
<td>-7485.45</td>
<td>15496.49</td>
<td>15122.89</td>
<td>15572.49</td>
</tr>
<tr>
<td>4-Pc Model</td>
<td>-7042.29</td>
<td>14859.15</td>
<td>14308.58</td>
<td>14971.15</td>
</tr>
<tr>
<td>5-Pc Model</td>
<td>-6804.69</td>
<td>14632.91</td>
<td>13905.38</td>
<td>14780.91</td>
</tr>
<tr>
<td>6-Pc Model</td>
<td>-6661.4</td>
<td>14595.29</td>
<td>13690.8</td>
<td>14779.29</td>
</tr>
<tr>
<td>7-Pc Model</td>
<td>-6541.04</td>
<td>14603.54</td>
<td>13522.08</td>
<td>14823.54</td>
</tr>
<tr>
<td>2-Sc 5-Pc Model</td>
<td>-6647.69</td>
<td>14332.74</td>
<td>13595.39</td>
<td>14482.74</td>
</tr>
<tr>
<td>2-Sc 6-Pc Model</td>
<td>-6488.3</td>
<td>14262.91</td>
<td>13348.59</td>
<td>14448.91</td>
</tr>
<tr>
<td>2-Sc 7-Pc Model</td>
<td>-6415.02</td>
<td>14365.33</td>
<td>13274.04</td>
<td>14587.33</td>
</tr>
<tr>
<td>3-Sc 5-Pc Model</td>
<td>-6613.01</td>
<td>14277.2</td>
<td>13530.01</td>
<td>14429.2</td>
</tr>
<tr>
<td>3-Sc 6-Pc Model (with covariates)*</td>
<td>-6479.56</td>
<td>14259.28</td>
<td>13335.13</td>
<td>14447.28</td>
</tr>
<tr>
<td>3-Sc 6-Pc Model (with covariates)</td>
<td>-6328.24</td>
<td>14039.62</td>
<td>13056.47</td>
<td>14239.62</td>
</tr>
<tr>
<td>3-Sc 7-Pc Model</td>
<td>-6407.54</td>
<td>14364.2</td>
<td>13263.08</td>
<td>14588.2</td>
</tr>
</tbody>
</table>

*Chosen model

Regardless of the number of classes, the scale-adjusted latent-class models were always preferred to standard preference-class models based on the information criteria. Adding scale covariates (observed factors which might influence choice consistency) improved the model fit further. The final model selected was a 3-Sc-6-Pc model with both preference and scale covariates. Table 4.11 describes parameters of the 3-Sc-6-Pc model, with only the preference and scale covariates which were statistically significant in at least one preference and scale class included. Preference parameters for Sc-2 and Sc-3 can be obtained by multiplying the preference parameters from Sc-1 by their scale factor.
Table 4.11: Preference-classes in scale-class one of Model 4

<table>
<thead>
<tr>
<th>Preference-class</th>
<th>Pc-1</th>
<th>Pc-2</th>
<th>Pc-3</th>
<th>Pc-4</th>
<th>Pc-5</th>
<th>Pc-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC (on none)</td>
<td>-171.997***</td>
<td>-153.514***</td>
<td>-9.971***</td>
<td>-18.993***</td>
<td>1.189</td>
<td>-6.159***</td>
</tr>
<tr>
<td></td>
<td>(65.932)</td>
<td>(53.775)</td>
<td>(2.253)</td>
<td>(3.869)</td>
<td>(1.366)</td>
<td>(1.437)</td>
</tr>
<tr>
<td>Detect</td>
<td>2.785***</td>
<td>0.846***</td>
<td>0.279**</td>
<td>0.152***</td>
<td>-0.104</td>
<td>0.956***</td>
</tr>
<tr>
<td></td>
<td>(0.518)</td>
<td>(0.184)</td>
<td>(0.110)</td>
<td>(0.055)</td>
<td>(0.268)</td>
<td>(0.288)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.331***</td>
<td>-0.999***</td>
<td>-0.504***</td>
<td>-0.155***</td>
<td>-0.723***</td>
<td>-1.704***</td>
</tr>
<tr>
<td></td>
<td>(0.083)</td>
<td>(0.234)</td>
<td>(0.140)</td>
<td>(0.041)</td>
<td>(0.234)</td>
<td>(0.421)</td>
</tr>
<tr>
<td>Cost</td>
<td>-1.003***</td>
<td>-0.303***</td>
<td>-5.354***</td>
<td>-2.268***</td>
<td>-2.417***</td>
<td>-0.638***</td>
</tr>
<tr>
<td></td>
<td>(0.229)</td>
<td>(0.086)</td>
<td>(1.208)</td>
<td>(0.478)</td>
<td>(0.770)</td>
<td>(0.173)</td>
</tr>
<tr>
<td>Pc proportions</td>
<td>32.35%</td>
<td>29.15%</td>
<td>7.48%</td>
<td>18.70%</td>
<td>7.73%</td>
<td>4.60%</td>
</tr>
<tr>
<td>Preference covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unconcerned dummy**</td>
<td>-0.337***</td>
<td>-0.202*</td>
<td>0.231</td>
<td>0.117</td>
<td>0.120</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td>(0.128)</td>
<td>(0.115)</td>
<td>(0.150)</td>
<td>(0.126)</td>
<td>(0.148)</td>
<td>(0.199)</td>
</tr>
<tr>
<td>Employed dummy*</td>
<td>0.190**</td>
<td>0.029</td>
<td>-0.183</td>
<td>0.127</td>
<td>0.095</td>
<td>-0.257*</td>
</tr>
<tr>
<td></td>
<td>(0.080)</td>
<td>(0.079)</td>
<td>(0.120)</td>
<td>(0.096)</td>
<td>(0.123)</td>
<td>(0.155)</td>
</tr>
<tr>
<td>White dummy**</td>
<td>0.227</td>
<td>0.289</td>
<td>0.048</td>
<td>-0.283</td>
<td>-0.508***</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td>(0.189)</td>
<td>(0.192)</td>
<td>(0.241)</td>
<td>(0.174)</td>
<td>(0.188)</td>
<td>(0.353)</td>
</tr>
<tr>
<td>IAP dummy</td>
<td>-0.064</td>
<td>0.111</td>
<td>0.054</td>
<td>0.025</td>
<td>-0.082</td>
<td>-0.045</td>
</tr>
<tr>
<td></td>
<td>(0.074)</td>
<td>(0.075)</td>
<td>(0.115)</td>
<td>(0.091)</td>
<td>(0.116)</td>
<td>(0.152)</td>
</tr>
<tr>
<td>Scale-class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale proportions</td>
<td>23.84%</td>
<td>64.87%</td>
<td>11.29%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale factor</td>
<td>1</td>
<td>0.139</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Failed internal validity dummy**</td>
<td>.</td>
<td>-0.088</td>
<td>6.630**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(4.133)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task difficulty***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very easy (1)</td>
<td>.</td>
<td>-0.815***</td>
<td>-3.358</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.292)</td>
<td>(2.225)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy (2)</td>
<td>.</td>
<td>-0.129</td>
<td>-3.072</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.312)</td>
<td>(2.349)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neither easy/ hard (3)</td>
<td>.</td>
<td>0.139</td>
<td>2.197**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.340)</td>
<td>(0.954)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hard (4)</td>
<td>.</td>
<td>0.759</td>
<td>1.606</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.490)</td>
<td>(1.175)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very hard (5)</td>
<td>.</td>
<td>0.046</td>
<td>2.627*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.958)</td>
<td>(1.496)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IAP dummy</td>
<td>.</td>
<td>0.042</td>
<td>0.220</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.106)</td>
<td>(0.305)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p<0.10; **p<0.05; ***p<0.01; standard errors in parentheses

Most women (80%) fell into either Pc-1, Pc-2 or Pc-4, which were the largest classes. Pc-3 accounted for 7.5% of the sample, Pc-5 for 7.7%, and Pc-6 for 4.6% Across all classes, coefficients for each attribute had the expected signs and were statistically significant (apart from detect in class five which was insignificant). Although a number of covariates were hypothesised to influence preference-class membership (education, religion, age,
children, experience of mammography, and experience of breast cancer with either friends and/or family), the only significant covariates were ‘unconcerned about risk of breast cancer’, ‘employment status’ and ‘ethnicity’. Figure 4.4 shows the relative importance of the attributes and the ASC for the opt-out for each preference-class.

**Figure 4.4: Relative importance of attributes by preference-class**

![Relative importance of attributes by preference-class](image)

The largest preference-class was Pc-1 which accounted for almost a third of the sample (32.3%). Women in this class treated the detection attribute as the most important, and had a large negative coefficient on the ASC for ‘No Screening’, suggesting a large amount of disutility in the absence of a breast screening programme. The women in this group were significantly less likely to be unconcerned about their risk of breast cancer (they had no idea, were concerned or were very concerned). The women in this class were also statistically significantly more likely to be employed.

Pc-2 was similar to class one. However, in this class, risk of unnecessary follow-up was more important than the attribute probability of detecting a cancer. Again, the coefficient on ‘No Screening’ was negative, large and highly significant suggesting women acquired utility from participating in the programme beyond that explained by the levels. Women in this class were also significantly likely to be unconcerned about their risk of breast cancer. This was a large class, accounting for 29% of the sample.

Another large class (18.7%) was Pc-4. In this class, the ASC on ‘No Screening’ was significant, negative and large (in line with the other preference-classes). However, women in this class were highly sensitive to cost, with this attribute being the most important.

In Pc-5, the ASC on ‘No Screening’ was positive (implying women in this group would rather not attend a screening programme) although statistically insignificant. This implies, in contrast to the other preference-classes, there was no baseline utility derived from participating in the screening programme. The most important attribute in this class was
cost, followed by risk of unnecessary follow-up. The coefficient on the detection attribute was negative and the smallest of all preference classes. The coefficient of the detection attribute was also not statistically significant suggesting it was not an important factor of screening choice for women in this preference class. To test if women in this preference class ignored the attribute (exhibited attribute non-attendance (ANA)), the model was restricted; constraining the coefficient on the detection parameter to zero. This definition and assessment of ANA was based on published quantitative examples (see Gibson et al., 2015; Lagarde, 2012; Scarpa et al., 2009) which rely on an ex-post measurement rather than inferring attendance directly from survey questions. Restricting the model to allow for non-attendance to the detection attribute in Pc-5 improved model fit (BIC reduced to 14023.53), suggesting women in this class completely ignored this attribute. These women were significantly more likely to be from ethnic minorities (non-White). There is existing literature which could explain why this group of women who choose not to attend screening, and this is drawn upon in the discussion presented in section 4.7.

The two smallest classes were Pc-3 and Pc-6. Pc-3 accounted for only 7.5% of the women sampled. In the smallest class, Pc-6, the coefficient on the ASC for ‘No Screening’ was relatively small (compared with other classes) although still statistically significant. Women in this class were statistically significantly less likely to be employed.

Relative to this base class (Sc-1), the scale factor in Sc-2 was 0.14 (high error variance) and in Sc-3 was 0.01 (very high error variance). Respondents were split between the three scale-classes with 23.8% in Sc-1; 64.9% in Sc-2; and 11.3% in Sc-3. Three-quarters of the respondents had a relatively high error variance (low choice consistency).

The largest scale-class was Sc-2 which accounted for almost 65% of the sample. The women in this class were significantly less likely to have reported the task as being ‘very easy’, which is in line with the high error variance and low choice consistency associated with a scale factor of 0.14. In Sc-3, the women were significantly more likely to have failed the internal validity test and report the task as hard, again in line with the very high error variance (low choice consistency) in this class.

Risk format was not a statistically significant covariate in either preference-class or scale-class membership suggesting the formatting of risk did not affect either respondents’ preferences or their choice consistency.

4.4.7 Marginal rates of substitution
To generate the MRS between the attributes, ratio calculations were conducted. The MRS explain the rate at which women were willing to give up the benefits of screening
(probability of detecting a cancer) in exchange for increased financial cost or increased risk of unnecessary follow-up, whilst maintain the same level of utility.

4.4.7.1 Willingness-to-Pay

The MWTP estimates from the heteroscedastic conditional logit for each class are shown in Table 4.12. The variation across the different preference-classes illustrated the heterogeneity in women’s preferences. The MWTP for a one percentage increase in the probability of detecting a cancer ranged from £5.20 in Pc-3 to £279.48 in Pc-2. Pc-1 and Pc-2 have large MWTP for detection relative to the other classes, and considerably larger than the heteroscedastic conditional logit results which were £79.17. The MWTP for a one percentage decrease in the risk of unnecessary follow-up ranged from £6.83 in Pc-4 to £330.15 in Pc-2. Pc-2 and Pc-6 have very large MWTP for the attribute ‘risk of unnecessary follow-up’, suggesting these classes were very intolerant of this risk.

In Pc-5, MWTP was calculated using the insignificant parameter estimates hence MWTP to detect a cancer is negative (indicating that women in this class would need to be compensated for better detection).

Using current estimates for the levels of the screening attributes, a total WTP for a breast screening programme was estimated in Table 4.13. These values were calculated using the estimated parameters from the heteroscedastic conditional logit model in column one of Table 4.13 and the estimated parameters from each preference class in the scale-adjusted latent-class model. The levels for the screening programmes were taken from Table 4.3. The results of the heteroscedastic conditional logit show that WTP for best-case screening scenario (with no risk of unnecessary follow-up and a 14% detection rate) was over three times higher than for the worst-case scenario (with 3% probability of detecting a cancer but a 20% risk of unnecessary follow-up).

The WTP for the screening programmes in Table 4.13 again highlights the preference heterogeneity within the sample. The large values in Pc-1 and Pc-2 are generated by the

<table>
<thead>
<tr>
<th>Table 4.12: MWTP for breast screening attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heteroscedastic conditional logit</td>
</tr>
<tr>
<td>MWTP detect</td>
</tr>
</tbody>
</table>

*Insignificant coefficient on detection; confidence intervals in parentheses
high values of the ASC on the ‘No Screening’ option, essentially reflecting that women in this group have a very strong preference to participate in a NHSBSP. In contrast, individuals in Pc-5 would pay nothing for screening in the best-case scenario, and they would have to be compensated to attend screening in the intermediate or worst-case scenarios. Similarly, whilst individuals in Pc-6 were willing to pay for the screening programme in the best-case scenario, they would have to be compensated for participation in the intermediate and worst-case scenarios.

Table 4.13: WTP for a lifetime breast screening programme

<table>
<thead>
<tr>
<th></th>
<th>Heteroscedastic conditional logit</th>
<th>Pc-1</th>
<th>Pc-2</th>
<th>Pc-3</th>
<th>Pc-4</th>
<th>Pc-5*</th>
<th>Pc-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening worst-case</td>
<td>£772.86 (£734.69-£812.32)</td>
<td>£17,329.25</td>
<td>£44,967.88</td>
<td>£13.42</td>
<td>-£597.97</td>
<td>-£3,925.49</td>
<td></td>
</tr>
<tr>
<td>Screening intermediate</td>
<td>£1,787.13 ( £1,714.75- £1,860.17)</td>
<td>£19,604.36</td>
<td>£50,225.05</td>
<td>£144.05</td>
<td>-£298.99</td>
<td>-£206.36</td>
<td></td>
</tr>
<tr>
<td>Screening best-case</td>
<td>£2,563.90 ( £2,493.87- £2,633.97)</td>
<td>£21,046.05</td>
<td>£54,643.89</td>
<td>£259.07</td>
<td>£931.64</td>
<td>£3,063.23</td>
<td></td>
</tr>
</tbody>
</table>

* Insignificant coefficient on detection; confidence intervals in parentheses

4.4.7.2 Risk tolerability

The WTA an increased risk of unnecessary follow-up for a one percent increase in the probability of detecting a cancer is shown in Table 4.14. The results showed that women were willing to accept nearly two additional unnecessary follow-ups (1.72) for an additional cancer detected, per 100 women screened. Again, results from the scale-adjusted latent class analysis indicate substantial preference heterogeneity with WTA ranging from 0.55 unnecessary follow-ups for an additional cancer detected in Pc-3, to 8.41 unnecessary follow-ups in Pc-1. In Pc-5, women were completely intolerant of the risk as they derived no utility from the detection attribute.

Table 4.14: WTA risk to detect an additional cancer

<table>
<thead>
<tr>
<th></th>
<th>Heteroscedastic conditional logit</th>
<th>Pc-1</th>
<th>Pc-2</th>
<th>Pc-3</th>
<th>Pc-4</th>
<th>Pc-5*</th>
<th>Pc-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willingness-to-accept risk</td>
<td>1.72 (1.47-1.97)</td>
<td>8.41</td>
<td>0.85</td>
<td>0.55</td>
<td>0.98</td>
<td>-0.14</td>
<td>0.56</td>
</tr>
</tbody>
</table>

* Insignificant coefficient on detection; confidence intervals in parentheses

Standard errors, or confidence intervals, were not included for the preference classes in Tables 4.12, 4.13 or 4.14 due to the challenges of exporting variance-covariance matrices from a scale-adjusted latent-class models with scale-covariates in LatentGold®. The variance-covariance matrix was required for simulation procedures, which produce
confidence intervals and standard errors, such as the Krinsky-Robb method (Krinsky & Robb, 1986).

4.4.7.3 Predicting uptake

Using responses from the heteroscedastic conditional logit, the probability of a woman participating in screening was estimated using equation 4.22. Assuming a probability of detecting a cancer rate of 10%; risk of unnecessary follow-up of 10%; and a lifetime screening cost of £100, the probability of a woman participating in screening is given by:

\[ P_l = \frac{e^{V_i}}{e^{V_i} + e^{V_j}} = e^{(\beta_1 10 + \beta_2 10 + \beta_3 1)} e^{(\beta_1 10 + \beta_2 10 + \beta_3 1) + e^a} = \frac{1.267}{1.491} = 0.850 \]  \[4.22\]

In this case, the probability of a woman choosing screening over ‘No Screening’ was 85%. Table 4.15 shows how this varied for the different screening scenarios described in Table 4.3. In the best-case scenario, uptake was predicted to be over 90%. In the worst, two-thirds of women would still attend. Only when risk of unnecessary follow-up was 47% (given detection at 10%) did the screening uptake drop to 50%. In a catastrophic scenario of no cancers being detected and all women in the programme receiving some unnecessary follow-up, 4% of women would still participate in screening.

<table>
<thead>
<tr>
<th>Screening worst-case</th>
<th>Screening best-case</th>
<th>Screening intermediate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of participating in screening</td>
<td>66.8% (66.47-66.80)</td>
<td>92.6% (92.57-92.72)</td>
</tr>
<tr>
<td>Confidence intervals in parentheses</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.5 shows graphically the effect increasing the risk of unnecessary follow-up on the probability of a woman participating in a screening programme for the different preference classes and the heteroscedastic conditional logit (hetclog) holding detection rates constant at 3% and cost constant at £20 per screen. For Pc-1 and Pc-2, women will always attend regardless of the risk of unnecessary follow-up. In Pc-4, women were very tolerant but uptake diminished as the risk of unnecessary follow-up exceeded 75%. In Pc-3 and Pc-6, women were tolerant of low levels of risk. In Pc-5, there was only a positive probability of a woman participating in screening at very low levels (<2%) of risk.
Figure 4.5: The effect of increasing risk on uptake for breast screening by preference-class

Figure 4.6 shows graphically the effect of increasing the probability of detecting a cancer on the probability of a woman participating in a screening programme holding risk of unnecessary follow-up constant at 5% and cost constant at £20 per screen. Women in Pc-1, Pc-2 and Pc-4 would always attend for screening, even at very low (and zero) levels of detection. In Pc-3, there were still high levels of uptake when probability of detecting a cancer is low and all would attend at detection rates higher than around 10%. Some women in Pc-5 would not attend for screening, regardless of the detection rate. In Pc-6, was low levels of uptake when the probability of detecting a cancer was low, but, similar to Pc-3, would always attend screening at detection rates higher than approximately 10%.

Figure 4.6: The effect of increasing detection rates on uptake for breast screening by preference-class

In Figure 4.7, the uptake rates given varying levels of screening programme cost is presented for the different preference classes and the heteroscedastic conditional logit
holding detection constant at 3% and risk constant at 5%. At zero cost, there was some uptake in all classes; however, this was very low in Pc-5 (where women were highly sensitive to the price of the screening programme). In Pc-3, Pc-4 and Pc-6, the probability in participating in screening dropped substantially as the lifetime cost of screening approached £1,000. Uptake rates in Pc-1 and Pc-2 were largely insensitive to cost even at values greater than £4,000.

**Figure 4.7: The effect of increasing cost on uptake for breast screening by preference-class**

4.5 Discussion

This study investigated whether the communication formats of risk attributes affected respondents’ choices (or utilities) in a DCE. The results provided confirmatory evidence that the DCE is theoretically valid, with many women trading the attributes presented and results showing significant coefficients with signs in line with *a priori* expectations. In addition, sample characteristics indicated successful randomisation of DCE respondents to each risk format.

The key findings will be discussed in section 4.7.1, with reference to the original research questions stated in section 4.2. The implications of these findings for this thesis and research more generally is presented in section 4.7.2, and in section 4.7.3, the results of this study are compared to the existing literature. The strengths and limitations of the study are discussed in section 4.7.4.

4.5.1 Key findings

The overarching aim of this research was to understand if, and how, the communication of risk attributes in a DCE influenced respondents’ choices, the consistency of these choices,
and the valuations derived. In this study, results from multiple models found no statistically significant difference in scale or preferences between the two risk formats.

The results of the heteroscedastic conditional logit model (Model 2A) found that the interaction terms between IAP and the risk and detection attributes were insignificant and that IAP had an insignificant effect on the scale term. This suggests that the risk format did not lead to any direct differences in either marginal utility or choice consistency, implying that the communication of risk had no effect in this study sample. These results were confirmed in the scale-adjusted-latent-class analysis (Model 4) which found that the risk format was not a significant predictor of either preference-class or scale-class membership. These results answer research questions two and three: randomness of choice and estimated utility parameters are not affected by the communication of risk.

This result is, in some respects, surprising because the structured review presented in Appendix 3.1 suggested risks were better understood when communicated with icon arrays. However, the review also identified that risk communication comprises multiple components, including risk perception and experience. In the latent-class models, concern about breast cancer risk was a statistically significant predictor of preference-class membership, indicating that risk perception, rather than risk communication, plays a significant role in women’s preferences and their valuations.

Research question four was investigated through latent-class analysis which found women’s preferences for breast screening were highly heterogeneous, with six distinct preference-classes. The largest class, class one, comprised women who tended to be more concerned about their risk of breast cancer and were significantly more likely to be employed. There was a high utility associated with participating in a programme with detection being the most important attribute in this class.

In Pc-5 five, women appeared to be uninterested in the screening programme with the estimate of the ASC and detection attribute both positive and insignificant. These women tended to be from ethnic minorities which is consistent with a large literature acknowledging the low up-take rates from these populations (Hoare, 1996; Renshaw et al., 2010), and that White women are most likely to attend screening, with little difference between age groups (Renshaw et al., 2010; Bansal et al., 2012). To target increased uptake for women from ethnic minorities, the benefits of screening could be communicated better. One possible option is to offer GP endorsement letters and multilingual leaflets, for which there is existing evidence that this improves uptake for ethnic minorities (Bell & Branston, 1999). Alternatively, women in this group could be paid (or receive a subsidy for
their incurred out-of-pocket costs) for screening. A study found paying for transport improved attendance by 16% in ethnic minority groups (Bell & Branston, 1999).

The statistically significant predictors of class membership were ethnicity, employment status and concern about risk of breast cancer; this answered research question five. Other characteristics of the women (such as age or education) were not significant predictors of class membership.

Research question six sought to understand the MRS between attributes. The MWTP of women in preference-classes two, three, four, five and six was higher for an additional risk of unnecessary follow-up than for an additional cancer detected. This implies that women place more weight on the risks associated with breast screening than may have been previously thought. However, in preference-class one, the largest class, MWTP to detect a cancer was greater than that to avoid risk of unnecessary follow-up. Similarly, in preference-class one, women were about eight times more tolerant of the risk of unnecessary follow-up than women in the other preference-classes. The MRS highlights the heterogeneity between the distinct preference-classes.

The results from the heteroscedastic conditional logit model gave a WTP of £1783 (about £350 per screen). Private screening costs around £250 (BreastHealthUK, 2014) and allowing for some selection bias (women more interested in breast cancer and therefore more likely to attend screening completing the DCE) this is comparable to the market rate. Again allowing for some selection into DCE completion, the percentage uptake of around 85% also approximates the true uptake rates in England, with recent estimates indicating these to be around 75% (Breast Cancer UK, 2015).

4.5.2 Key implications

Three key implications for researchers conducting DCEs to elicit preferences for benefit-risk trade-offs were found: 1) pictorial framing of the risk attribute in addition to the percentage made no difference to either marginal utility or choice consistency; 2) risk preferences, in the case of screening for breast cancer, were highly heterogeneous; and 3) accounting for scale heterogeneity improved model fit.

What affects the valuation of risk in a benefit-risk trade-off DCE may not necessarily be due to the communication format of the attribute. An important factor affecting respondents’ choices and valuation could be their concern, or perception, about their own risk. Research conducting DCEs for benefit-risk trade-offs should consider the sample’s risk perception through additional survey questions and acknowledge the impact of this in their derived valuations.
There was significant preference heterogeneity in the model and the modelling data to allow for different preference-classes improved fit over a one-class heteroscedastic conditional logit model (BIC 19877.922 v 14595.29). Using latent-class models can also help understand preferences, when the drivers of heterogeneity are latent (or unknown). This is in line with other research which has found modelling preference heterogeneity with latent models (as opposed to interacted attributes and socio-demographic variables) is preferred, particularly in a healthcare context (Hole, 2008).

The study also highlighted the importance of accounting for scale heterogeneity in respondents’ choices. Self-reported task difficulty was a significant predictor of scale-class membership. There is already existing evidence to suggest scale heterogeneity is a major issue, particularly in healthcare DCEs (Fiebig et al., 2010; Flynn et al., 2010). Allowing taste and scale heterogeneity to be confounded may result in biased estimates (Adamowicz et al., 2008), and if scale is not accounted for then apparent differences in estimates across groups may not be due to true differences in preferences.

4.5.3 Comparisons with existing literature

This section compares different relevant literatures with the results and key findings of the empirical study. In section 4.5.3.1, the literature surrounding women’s preferences for the benefits and risks associated with a NHSBSP are explored and contrasted to the results of this study. In section 4.5.3.2, the findings relating to the risk communication format are contrasted with evidence from the review of Appendix 3.1. In section 4.5.3.3, other DCEs eliciting preferences for benefit-risk trade-offs are drawn upon to explore the robustness of the quantitative methods used in this study.

4.5.3.1 Literature on screening preferences

There have been no comparable DCEs looking at women’s preferences for the benefit-risk trade-offs associated with screening for breast cancer (as identified in Chapter Three, section 3.4.6). However, a qualitative study investigating Australian women’s views on over-diagnosis found that “the lower and intermediate estimates [of over-diagnosis] (1–10% and 30%) had limited impact on attitudes and intentions, with many women remaining committed to screening” (p.1; Hersch et al., 2013). Whilst this is in line with the results of the heteroscedastic conditional logit model (Model 2), the latent-class analysis did indicate substantial heterogeneity around these views. The findings from Hersch et al., (2013) do not reflect those of other large preference groups; at a 30% risk of unnecessary follow-up uptake would be very low (almost zero) in preference-classes three, five and six (accounting for around a fifth of women sampled). The quantitative exploration of
heterogeneity in women’s preferences for a NHSBSP presents a contribution to the literature on screening preferences.

4.5.3.2 Risk communication literature

In the rapid review of risk communication studies presented in Appendix 3.1. Icon arrays were chosen as a risk communication format with the most empirical support, although one study communicating breast cancer risk (Brewer et al. (2012)) found that icon arrays were the least preferred method with 39% (n=46) of women saying it was risk format they liked the least (risk ladder was most preferred). Therefore, although the format had not been tested in a choice or trade-off setting, it was anticipated that IAP would be superior to PO. The hypothesis that IAP would improve choice consistency was not accepted. In addition, no differences were found between the utility parameters.

4.5.3.3 Studies considering framing effects in DCEs

There are some example DCE studies which have investigated the effect of different attribute frames. For example, Howard & Salkeld (2009) investigated the framing of risk attributes in a DCE for colorectal cancer screening. The DCE contained four framing formats for a risk attribute (true-positive/false-negative/true-negative/false-positive) in a forced choice experiment. The authors used a MXL and a MNL, comparing models on McFadden’s pseudo $R^2$ and AIC. The authors interacted framing method with each attribute to identify differences due to the frame and compared the coefficients of the risk attribute by looking for overlap in the standard deviations from the four versions (no overlap indicated framing significantly influenced valuation). However, Howard & Salkeld (2009) made no account for scale heterogeneity between the survey versions, so their conclusion that the framing of risk did influence WTP (contrary to the results of this thesis) may be driven by differences in error variance rather than marginal utility.

In a conference paper, Buchanan et al., (2014) trialled a forced choice DCE with two risk communication formats (frequencies and icon arrays) with a mixture of postal, internet, clinic-based surveys completed by patients in NHS trusts or members of a charity. The authors then ran four models (MNL; a split sample MNL; MXL; and a latent-class model) on the pooled data, and found that respondents receiving the icon array had higher WTP values (by approximately 50%). However, these authors again failed to account for scale heterogeneity between risk formats (and survey modes), directly comparing coefficients despite the confounding of scale and preferences.
The absence of consideration of the scale parameter in studies comparing risk formats limits the ability to contrast the results of this study. Other studies comparing the effect of cost attribute framing have also failed to account for differences in scale heterogeneity (for example, Grutters et al. (2008)). It appears to be a consistent limitation in healthcare DCE analysis, with applied studies not accounting for sources of bias (most notably the scale parameter) identified in other research areas such as marketing and environmental economics (Louviere, 2006; Fiebig et al., 2010).

4.5.4 Strengths and limitations

As section 4.5.3.3 showed, this was an original empirical investigation, going beyond current standard of analysis in typical health DCEs. The application was also original in that there are no published examples comparing percentages and icon arrays in the framing of risk attributes. The DCE also possesses characteristics of both theoretical and external validity, with results in line with a priori expectations and comparable to revealed preference data.

However, the study had some limitations. A heteroscedastic mixed logit model could have been used as a further investigation into preference heterogeneity. However, it was believed that describing heterogeneity in terms of latent classes (rather than say a distribution from a mixed logit model) would be more useful for decision-makers. In addition, latent class models are arguably superior to mixed logit models in their statistical properties (Shen, 2009).

Non-linearity in preferences was investigated in the preliminary stages of model selection (see Model 1D). Investigations of non-linearity could not be conducted in the more complex models, such as the selected 3-Sc 6-Pc Model as this failed to converge. Non-convergence is likely to have occurred due to the increased number of parameters (from three to 14) and the small sample size in some preference classes, such as preference class six which had fewer than 50 respondents. However, there is a possibility that some, or all, preference classes may have non-linear preferences in some or all attributes, and this should be considered when interpreting the results presented in Table 4.11.

The conclusion that IAP offered no advantage to respondents in terms of reduced cognitive burden and had no effect on their preferences, may be specific to the context of this study (breast screening) or the magnitude of the levels used (percentages were whole numbers). In other scenarios, the results may differ, and further research is required to investigate the generalisability of the key findings described in section 4.7.1. Similarly, alternative
methods (such as decimals, frequencies or fractions) may differ to percentages and icon arrays.

Furthermore, the quantitative analysis was based on observed data and unobserved behaviour (for example, what respondents were truly thinking) was not collected. It could be that this was affected by the framing of risk even if choice behaviour was ultimately the same.

4.6 Conclusion
This study has helped to understand more about the effect of different risk communication formats. It was identified that neither choice nor choice consistency were affected by the framing of the risk attributes, but that there was significant preference heterogeneity, with latent-classes of women deriving different values from different aspects of a breast screening programme. However, the analysis relied on a purely quantitative interpretation of women’s trading behaviour.

In order to investigate risk communication effects further, alternative research methods may be required to understand more about women’s interpretation of the attributes in a DCE context. The following chapter uses a qualitative research method called think-aloud to further investigate framing effects.
Chapter Five

Benefit-risk trade-offs for breast screening: a qualitative study

5.1 Introduction

This chapter describes a study designed to explore more about how respondents make choices in DCEs containing risk attributes. The chapter uses qualitative research methods as part of a mixed-methods approach to complement the purely quantitative approach of the previous study presented in Chapter Four. Qualitative research methods were used to understand more about how the format of risk might affect respondents’ cognitive heuristics in a DCE setting and explore unobservable factors which might influence choice-making behaviour.

Chapter Five begins in section 5.2 by describing a systematic review of qualitative research alongside published DCEs and a survey to the authors of these DCEs eliciting their views and experiences of the usefulness of such methods. The research questions to be investigated are described in section 5.3. A description of the types of qualitative research methods available are presented in section 5.4, before a justification and explanation of the methods and analysis used in this study are presented in section 5.5. In section 5.6, the results are presented using the key concepts: risk and decision-making strategies. These concepts were identified through a literature review determining these a priori as important areas in addressing the overall aim of the PhD, which is to understand the drivers and consequences of risk communication format in a DCE most effectively. Chapter Five concludes with a discussion and description of the limitations of this study in section 5.7, before briefly introducing how the subsequent chapter (Chapter Six), an eye-tracking study, seeks to address some of these concerns in section 5.8.

5.1.1 Background to qualitative research methods

Qualitative research methods are used to understand what, how and why through explorations of people’s views or lived experiences (Silverman, 2013). Chapter Four identified substantial heterogeneity in women’s preferences for a NHSBSP. The quantitative analysis was limited in its ability to answer questions about what people think when they see the risk attribute, how they compare risks and benefits, and why they make such choices. In order to better understand how people balance risks and benefits, and to acquire a deeper understanding of DCE respondents’ valuation strategies, qualitative research methods were identified as an appropriate strategy.
5.2 Qualitative research methods and DCEs

While DCEs involve statistical analysis of quantitative choice data, other research areas, particularly in the social sciences, have used qualitative research methods to investigate people’s beliefs and decision-making processes (Berg, 2007). Although not frequently employed by economists, there is a growing body of research supporting the use of qualitative methods in health economics (Coast, 1999; Coast et al., 2004). In the context of DCEs, qualitative research methods have a number of potential roles. This could include identifying attributes and/or their associated levels, refining the terminology, cognitive piloting of the survey, exploring the use of aids and vignettes, gaining an understanding of respondents’ decision-making processes, or estimating preference heterogeneity before modelling the data (Coast & Horrocks, 2007; de Bekker-Grob et al., 2012; Kløjgaard et al., 2012).

Guidelines for conducting DCEs advocate qualitative methods with specific recommendations for the development of attributes and levels (Bridges et al., 2011b; Coast & Horrocks, 2007; Lancsar & Louviere, 2008). However, there has been no explicit investigation of how well these recommendations regarding the use of qualitative methods in the preparation and testing of DCEs have been translated into practice.

5.2.1 Systematic review of qualitative research methods alongside DCEs

5.2.1.1 Systematic review aims

The aim of this component of the thesis was to systematically identify all published studies in healthcare that reported the use of qualitative methods to inform the design and/or the interpretation of DCEs. The objectives were to: summarise the proportion of DCEs using qualitative methods; assess the context in which the research was applied; identify the methods and techniques used; and, where possible, appraise the quality of the research conducted.

5.2.1.2 Systematic review methods

This study used systematic review methods as advised by the Centre for Reviews and Dissemination (CRD) to identify all healthcare DCEs published in the last decade (since, and including, a previous systematic review) (de Bekker-Grob et al., 2012). The systematic review focussed on identifying DCEs rather than other stated preference methods such as CJA, ACA or contingent valuation because, as discussed in Chapter Two, these methods are grounded in different economic theories and are therefore not directly relevant to this review or the overall thesis (Louviere et al., 2010).
An electronic search of Medline (Ovid, 1966 to date) was conducted in June 2012. Although other databases could have been searched, the strategy exactly replicated that of published reviews of DCEs (Ryan & Gerard, 2003b; de Bekker-Grob et al., 2012). The search terms used were: ‘discrete choice experiment(s)’, ‘discrete choice modeling’, ‘stated preference’, ‘part-worth utilities’, ‘functional measurement’, ‘paired comparisons’, ‘pairwise choices’, ‘conjoint analysis’, ‘conjoint measurement’, ‘conjoint studies’, and ‘conjoint choice experiment(s)’. The term ‘conjoint analysis’ was included to identify studies which had used discrete choices rather than those which required respondents to rate or rank alternatives. No search terms were used to directly identify qualitative studies as this was deemed to be too restrictive.

The main foci of the review were to: 1) identify and quantify the proportion of DCEs using qualitative methods; 2) investigate the stages in the DCE at which qualitative research is employed; 3) understand the methods and techniques currently used; and 4) where possible, evaluate the quality of the reporting of research. The studies were initially categorised into three categories: 1) those which reported no qualitative research; 2) those which contained basic reporting which indicated some qualitative research may have been used; and 3) those which indicated an extensive qualitative component was conducted in direct relation to the DCE. This categorisation identified studies, in category three, which contained sufficient detail for critical appraisal.

Using pre-existing guides for appraisal of qualitative research methods in a traditional sense may have meant the extensive studies identified by this review would have been judged incorrectly or unfairly. It is, however, crucial that the qualitative research contained in the studies was formally assessed in a standardised and systematic way. Therefore a bespoke appraisal tool was developed to include broader issues which were not included in the traditional tools as advised by the CRD (CRD, 2008). The bespoke appraisal tool can be found in Appendix 5.5.

5.2.1.3 Systematic review results

One hundred and twenty four studies were already identified by a previous systematic review (de Bekker-Grob et al., 2012). The search resulted in 501 titles and abstracts since the previous review (2008 onwards) and 208 full papers were retrieved for further assessment and 148 papers met the inclusion criteria. Figure 3.1 shows the stages involved in screening and the reasons for rejection of the excluded papers using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) review guide (Moher et al., 2009). Therefore, in total, there were 277 papers included in the final review, which related to 254 empirical studies (because some studies were reported in more than
one paper). A detailed description of the reasons for rejection can be found in Appendix 5.3. Overall, 111 (44%) studies did not report the use of any qualitative research methods; 114 (45%) reported minimal data on the use of qualitative methods; and 29 (11%) reported, or explicitly cited, the extensive use of qualitative methods.

Almost all (n=113, 99%) of the studies which reported basic qualitative research reported using it before the DCE was implemented, either in the design or piloting phase. A variety of applications of qualitative research methods were identified. In the design of the DCE, researchers were most commonly seeking to identify attributes and/or assign levels (n=70, 61%), or validate attributes and/or levels identified through other methods (n=31, 27%). Researchers also used qualitative research methods more specifically to check terminology, vignettes and descriptions (n=9, 8%) and to confirm translations (n=2, 2%). After the design phase, some studies also reported using qualitative research methods in the piloting of the DCE (n=24, 21%). In the pilot stage, the methods were specifically used to check for decision strategies (n=1, 1%) and also to determine an appropriate sample for the final DCE.

Of the studies reporting extensive details about the qualitative components, most studies (n=25, 86%) again reported the use of qualitative research methods to identify attributes and/or levels for use in the DCE. The most common data collection approach was interviews (n=18, 62%). These interviews were mostly semi-structured (n=12, 41%) and face-to-face, although two studies used telephone interviews. A number of studies also used focus groups or group interviews (n=12, 41%). Three studies (Cheraghi-Sohi et al., 2007; Ryan et al., 2009; Bridges et al., 2011b) used qualitative research methods to understand more about how respondents completed the choice task presented, and two of these (Cheraghi-Sohi et al., 2007; Ryan et al., 2009) used a type of verbal protocol analysis.

In terms of the analysis, most studies simply stated that they used thematic analysis (n=10, 34%) or content analysis (n=5, 17%) to categorise the qualitative data collected, with reducing the qualitative data to develop a few attributes and levels. Other analytical approaches included framework analysis (n=3, 10%) and a related approach to qualitative analysis called charting. Seven studies used some constant comparative analysis (n=4, 14%) or open-coding (n=3, 10%) at least in the initial stages.

5.2.2 A survey to authors

The systematic review presented in section 3.2 of this chapter was constrained by its reliance on the details reported in published academic papers. It is well established that a
bias exists in both medicine and social science, with studies exhibiting positive or interesting results being more likely to be published (Franco et al., 2014; Easterbrook et al., 1991) A survey was, therefore, designed to elicit information from authors of published DCEs in health about their views on the use of qualitative research methods alongside quantitative analytical methods.

5.2.2.1 Survey aims

The aim of this study was to try and reduce the effects of reporting bias through directly eliciting researchers’ views and experience of conducting qualitative research alongside DCEs in a healthcare setting. The objectives of this study were to reveal more about researchers’ experience of using qualitative research methods; their opinion of the usefulness of qualitative approaches in this context; and any explanation they had for the poor level of reporting found in the systematic review.

5.2.2.2 Survey methods

This study used an online semi-structured survey comprising closed and open questions. The survey was sent to authors who indicated, in a published study, that they had used qualitative research methods but only reported basic details. As described in section 5.2.1.3, total of 114 studies reported basic use of qualitative research methods and all authors were included in this review. As some corresponding authors had multiple studies included in the review, 91 individual authors were sampled. All of the journal articles provided an email address for the corresponding author. Therefore, the most feasible method of contacting authors and eliciting their views was to use an online survey.

The final survey (presented in Appendix 5.2) consisted of 12 questions and comprised a mix of closed-ended and free-text comment boxes. The questions asked authors about their experience and their opinion of: 1) using qualitative research methods alongside DCEs; and 2) communicating the qualitative work they conducted in a journal article. Additional questions included self-assessment of their and co-authors’ expertise in qualitative research, the number of DCEs they had conducted, and whether they agreed with the key finding of the systematic review that qualitative research is not well reported in healthcare DCEs.

The survey data were downloaded from the online server and analysed in Excel® (Microsoft, 2010). The analysis involved simple production of descriptive statistics for each of the questions. The authors’ free-text comments were not thematically analysed because of the limited textual data available (some authors chose not to comment).
5.2.2.3 Survey results

After the first email sent on 1\textsuperscript{st} May 2013, 38 authors completed the survey (an initial response rate of 42\%) within a month and before a second reminder email was sent. The questionnaire closed on the 30\textsuperscript{th} June 2013, with a total of 53 completed or partially complete responses, resulting in an overall response rate of 58\%.

The authors reported that the use of qualitative research methods added value to their experience of conducting DCEs in general, with 74\% (n=31) stating it made a ‘substantial improvement’ to the study. The majority of survey respondents (n=42, 79\%) agreed with the systematic review finding that the qualitative component was only briefly described in DCE papers. Half of the respondents (n=26, 50\%) believed the amount of qualitative work conducted was accurately reflected in the published paper as reported. Over half of the respondents (n=26, 52\%) considered qualitative research was too complicated to report in detail and one fifth of the respondents (n=10, 20\%) reported that they felt it was too time consuming to conduct properly. Some authors (n=11, 16\%) also stated that qualitative research would not be of interest to their peers, and n=4 (8\%) also reported that they did not believe it was important to funders. Appendix 13 tabulates the authors’ responses to each survey question.

5.2.3 Discussion of review and survey findings

The results of the systematic review and survey to authors identified qualitative research methods were being used by DCE researchers to answer multiple research questions, and that these methods add value to a DCE study. This finding suggested that qualitative research would be an appropriate method to use in this thesis.

The systematic review identified two papers (Cheraghi-Sohi et al., 2007; Ryan et al., 2009) which used concurrent verbal protocol analysis (called ‘think-aloud’) to understand more about people’s choices. As this PhD thesis seeks to understand more about how people answer a DCE containing a risk attribute, the think-aloud method was considered appropriate to answer the key research questions.

There was a paucity of detail about the analysis of the qualitative data collected, however, when described, most studies used some sort of thematic analysis, which was most often a pre-defined framework to code the collected data from which themes were developed using NVivo® software. Therefore the analysis of the think-aloud data collected in this thesis followed a similar approach.
5.3 Research questions

The key research questions addressed in Chapter Five were formulated from the findings of the systematic review (section 5.2 and Appendix 5.1) and the results of the quantitative study (see Chapter Four). The key research questions addressed in this chapter were:

1) What do people say they think about when they complete a DCE?
2) How do people complete DCEs with risk attributes?
3) How do people interpret risk attributes in a DCE?
4) Does the communication format of risk attributes affect DCE respondents’ qualitative accounts of choice making?

5.4 Introduction to qualitative research methods

Social science, and the research conducted by social scientists in particular, has been broadly split into two schools: positivism (which tests correlation between variables) and interpretivism (which is concerned with observation and description) (Silverman, 1993). Another way of distinguishing between these two schools of thought is whether hypothesis validation or hypothesis generation is sought (Johnson & Onwuegbuzie, 2004). For example, in generating a hypothesis, qualitative research methods could be used to begin an enquiry into a particular topic and quantitative research methods for confirmation of the topic’s existence. The exact distinction between the two schools of thought depends on the epistemological (founding beliefs) and ontological (philosophies) view taken by the researcher. Table 5.1 briefly outlines some of the differences between the two schools of thought.

Table 5.1: Schools of thought in research methodology

<table>
<thead>
<tr>
<th>School</th>
<th>Implied Hypothesis</th>
<th>Definition</th>
<th>Example of research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positivism</td>
<td>Hypothesis validation</td>
<td>Testing of correlation between variables by objective means and statistics</td>
<td>Confirming a relationship between variables; investigating generalizability</td>
</tr>
<tr>
<td>Interpretism</td>
<td>Hypothesis generation</td>
<td>Observing and describing by subjective means and points of view in an exploratory manner</td>
<td>Beginning a particular enquiry into a research topic: investigating transferability</td>
</tr>
</tbody>
</table>

Although contemporary economics can be seen as a mathematical-based science, and hence uses quantitative research methods with a positivist perspective, some have argued that all variables have some element of social construction (Silverman, 1993). Qualitative research methods largely fit under the interpretivist paradigm, designed to describe and
illuminate a particular topic, and thus offers the theoretical foundation to this study (Denzin, 2009).

Qualitative research methods collect and analyse textual data (in contrast to quantitative studies which traditionally seek numerical observations to test hypotheses) (Pope & Mays, 2008). The aim of qualitative research methods is to acquire a deeper understanding about a human action in reference to a particular topic or phenomenon (Huston & Rowan, 1998). To do this, qualitative research methods commonly employ interaction with people, and mostly use interviews or focus groups to initiate this interaction (Denzin & Lincoln, 2005).

Chapter Four empirically demonstrated that women’s preferences were highly heterogeneous but that the communication format of the risk attributes did not affect these preferences or the consistency of their choices. However, the large quantitative study and analysis was limited in its ability to explain more about how risk is interpreted by DCE respondents. The econometric modelling allowed identification of groups of women answering in similar ways, as there was evidence strongly suggesting the presence of distinct (latent) classes. Although it was possible to predict class membership (with some significant preference class co-variates), this interview study builds upon the findings of the quantitative analysis by exploring quantitatively, the unobservable drivers of choice and behaviour in more depth by directly asking respondents about their thoughts.

Furthermore, in-depth analysis is warranted as it is impossible to be certain that those surveyed all uniformly interpreted both the choice tasks and the terminology used. In other words, differences in preference classes may reflect differences in women’s interpretation rather than any actual differences in their attitudes or valuations. That is why qualitative research methods will be useful in revealing if, and how, respondents bring different meaning to the information presented to them.

The systematic review of the use of qualitative research methods in DCEs (see section 5.2 and Appendix 5.1) supported the use of qualitative methods to design and interpret DCEs. These uses are described in guidelines and in best-practice literature with the aim of improving DCE design and understanding more about DCE choices. The systematic review concluded that the design and analysis of qualitative research studies alongside DCEs were not well-reported. The supplementary survey to authors experienced in measuring peoples’ preferences confirmed that qualitative research methods were being used and, more importantly, in the experience of those surveyed, did have a positive influence on study progression and quality. The few DCE studies that reported the use of qualitative research in more detail, complemented by the results of the survey to DCE
authors, illustrated the potential power of qualitative research methods to reveal influences on choice or decision strategies that may be difficult, or impossible, to measure in a traditional quantitative DCE survey alone. These results from the review of section 5.2 provide a pragmatic justification for the use of qualitative research methods whose prime purpose are to understand what, how and why, from an individual perspective.

In this study, the qualitative approach involved the collection of empirical data to explore, in-depth, the views of DCE respondents. From these data, it was possible to understand more about how risk was interpreted in a DCE setting and if this differed by the risk communication method. The qualitative study was built around the DCE task and the choice sets produced by the design described in Chapter Three. The qualitative data provided a snapshot of the effect of completing the DCE under an interview setting.

What makes good quality qualitative research, particularly for stated preference studies, was discussed in the development of the bespoke critical appraisal tool employed in the systematic review of section 5.2. The bespoke critical appraisal tool served a second purpose in this chapter, Chapter Five, by providing a checklist of steps and structure for the design of the proposed qualitative study. The tool (see Appendix 5.5) can be summarised using the following headings: 1) background and overview; 2) description of the context; 3) sampling methods; 4) data collection; 5) reflexivity (or a consideration of how the investigators themselves can influence qualitative research); 6) ethical aspects; 7) data analysis; and 8) presentation of results. These headings will serve as a structure for sections 5.5 and 5.6 of Chapter Five which detail the methods and results, respectively.

5.5 Methods
The previous section, section 5.4, introduced qualitative research as a research method. The techniques used in qualitative research are now explored and critiqued. Following this, an explanation and description of the methods used in the empirical study are presented.

5.5.1 Qualitative research methods
The aim of qualitative research methods is to generate a textual account of an event from the perspective of the people being studied (Silverman, 1993). The most commonly used qualitative method is interviews. However, qualitative research methods include other approaches ranging from raw observations in the form of anthropological-enthographies (which involve no interaction with the subject of interest) through to the analysis of free-text comments or social media updates (Branthwaite & Patterson, 2011; Reeves et al., 2008).
Interviews are broadly defined by Burgess (1984) as: “conversation with a purpose”. They are not only the most popular method, interviews are also defined as the ‘gold standard’ of qualitative research methods (Silverman, 2013). Interviews are not one single method, rather a group of conversational approaches, which vary in the mode (phone, face-to-face, group settings) to the uniformity of the questions (structured, semi-structured, open-ended) (Denzin & Lincoln, 2005). Figure 5.1 summarises the possible research methods with the broad arrows containing possible conformations which could be employed with each approach.

**Figure 5.1: Conversational qualitative research methods**

Within the groups of conversational (interview) approaches, there are many different techniques which can be employed to conduct interviews. Possible approaches include group interviews, telephone calls, face-to-face settings or emails. The results of the systematic review presented in section 5.2 and Appendix 5.1 identified a verbal protocol method called concurrent think-aloud. This method has been used by a number of researchers (Cheraghi-Sohi et al., 2007; Ryan et al., 2009) specifically to understand more about people’s choice task performance (rather than other uses such as attribute identification or level development). For example, Ryan et al., (2009) looked specifically at how respondents processed the DCE task and what led to ‘irrational’ responses in dominant choice sets.

Whilst telephone interviews or email exchanges could have improved the response rate (and saved money and time) it would not have been practical in this study which used a
survey. Face-to-face interviews are generally seen as the gold standard (Novick, 2008) and can be particularly useful in a survey context where seeing the interviewee pointing, their body language, and facial expressions may yield additional insights into their thoughts (Mason, 2002). Although these were not analysed in this study, their observation to aid the interview process and provide a stimulus for the appropriate timeliness of prompts was deemed crucial. Therefore, given the limitations of other techniques and the success of the verbal protocol analysis in previous DCEs, concurrent think-aloud was identified as the most suitable qualitative research method for this study.

Development of the think-aloud method is attributed to Ericson & Simon (1984) who introduced the method to generate scientific data in psychological experiments investigating information retention in people’s short-term and long-term memory. Since then, the method has been widely used by both cognitive psychologists and interviewers seeking to understand people’s approaches to problem-solving (Boren & Ramey, 2000). Unlike other interviewing techniques, and contrary to the definition above, the think-aloud method does not involve much ‘conversation’. Instead, the dialogue is replaced by the interviewee speaking out-loud their thought processes with the interviewer prompting in periods of silence.

Semi-structured interviews lie in the midst of the qualitative methods spectrum, as they allow the interviewer to explore and probe areas of information that arise if the interviewee volunteers a topic; however, the presence of a schedule can assist in keeping the interviews structured enough to answer, rather than specifically develop, a research question. This study aimed to acquire understanding about people’s experience of completing DCEs and, therefore, a series of pre-determined topics that needed to be covered were identified. These pre-determined topics meant that a semi-structured interview in the form of debriefing questions was deemed appropriate in addition to the think-aloud protocol.

5.5.1.1 Sampling

Sampling is a crucial stage of any study design but is often constrained by resources (time and money), access to potential participants, and ethical considerations (such as the sensitive nature of the interview topic). Although commonplace in positivist research, probability-based sampling which aims for an unbiased random sample, is not a requirement of qualitative research (Marshall, 1996; Luborsky & Rubinstein, 1995). This study did not seek to produce a statistically representative sample, as statistical significance of any kind was, and is not, a goal of qualitative research. Instead a trustworthy sample was desired based on the transferability criteria recommend by Lincoln.
& Guba (1985). Sufficient detail about the final sample, fieldwork, context and environment were provided to enable readers of this study to evaluate the similarity of another setting and decide for themselves whether the findings were transferable. Therefore a non-probability purposive sampling strategy which aimed to collect a sample of diverse women was devised. In contrast to many qualitative research studies, which rely on principles of inductive reasoning to construct an idea (Burns, 1989), the primary aim was not to achieve hypothesis generation, but to understand how risk was interpreted and traded-off in the DCEs. Although a more sophisticated sampling strategy could have been used, a specific sample was neither desired nor sought.

The sample selection criteria were limited to females, fluent in English and between 18 and 70 years of age for reasons discussed in Chapter Three. Seventy years was chosen as a maximum as this is the current cut-off for routine screening in England (Independent UK Panel on Breast Cancer Screening, 2012), meaning women over 70 years old would be neither current users nor potential users. No other sample restrictions were included as initial discussions with breast cancer experts in the design of the DCE revealed that they were interested in the views of all women (see Chapter Three). In light of this, the internet panel (see Chapter Four) surveyed women in the 18 to 70 year age category and therefore, for consistency, the sampling frame for this qualitative study maintained the same inclusion criteria.

In line with purposive theoretical sampling, advertisements were placed in an attempt to acquire a range of interviewees with a range of key characteristics covering various ages, jobs and ethnicities. The advertising strategy involved placing advertisements online at The University of Manchester’s publically available volunteers’ website (www.student.manchester.ac.uk/volunteers) and through emails to staff distribution lists. Paper advertisements were placed in shops, cafes, bars and public noticeboards. Examples of these advertisements can be found in Appendices 3.9, 5.14 and 5.15. The advertisement responders were only incentivised at first contact with the researcher, when they were informed about the £10 Amazon voucher, as per the recommendation of the ethics committee. At this first point of researcher contact, the potential participants were also provided with an information sheet containing frequently asked questions (Appendix 3.11). The advertisement responders were then asked to read the sheet, take time to think about the study and then, if they were interested, make contact again to arrange the interview.
5.5.1.2 Interview schedule

A preliminary interview schedule was used in pilot interviews (n=5) described in Chapter Three. After these interviews were completed, it was decided that warm-up questions were required to encourage interviewees to vocalise their thoughts. The final interview schedule (Appendix 5.16) was created through an iterative process of discussions with the supervisory team and comprised a warm-up exercise and debriefing questions. The main part of the interview involved prompting the participant to keep verbalising their thoughts through encouraging questions (Why did you choose that? What are you thinking?) and did not follow a pre-defined structure. This also limited response acquiescence or a ‘yea-saying bias’ by provoking interviewees to come-up with their own accounts instead of agreeing with the interviewer (Blamey et al., 1999).

The interview schedule helped to keep the interviews on topic, and although useful to examine participants’ perspectives and perceptions of risk (including their family history and how this had influenced their responses) the structure allowed the interview to be brought back to topic whilst permitting exploration of emerging issues which were not anticipated and consequently not included in the guide. As discovered when developing the critical appraisal tool, care needs to be taken when applying qualitative research in a stated preference setting to ensure the questions do not lead to a pre-assigned idea or hypothesis and, instead, remain exploratory in nature (the so-called ‘tourism effect’ by Silverman (1993)). Similarly, there were no ‘right answers’ or a priori expectations to this research, and the flexible schedule allowed complete collection of (even disconfirming) data.

5.5.1.3 Data collection

Women who agreed to the interview were invited to either come to the university or meet in a location of their choice, as long as it was quiet and private enough to facilitate audio-recording. The participants were also reassured that they needed no prior subject knowledge. All participant contact and interviews were conducted by CV. Before the interviews started, interviewees were shown the information sheet they first received when they enquired about the study (Appendix 3.11) and the purpose of the study was reiterated. After the researcher had checked whether the interviewee had any questions, the interviewee completed a consent form permitting the use of recording and note-taking (Appendix 3.12). The digital voice recorder was then set to record and the interview began.

The interview started by explaining to the participant what was meant by think-aloud and that they did not need to talk to the interviewer per se, but just say whatever they were thinking. As identified to be useful in piloting, a warm-up exercise was conducted which
asked interviewees: “How many windows do you have in your house?”. If women gave an absolute number straight away, the think-aloud concept was explained again and they were invited to reveal how they got to their answer, which usually resulted in a walk-through of their home. Interviewees were then reassured that this was all they would be required to do and were then invited to read over the training material for the DCE to familiarise themselves with the topic and task to be completed.

Interviewees then completed the DCE, thinking aloud whilst selecting their answers on an iPad. Women were randomised to a risk communication format and experimental design, consistent with the method used in Chapter Four. At the end of the DCE, a series of debriefing questions was posed to elicit anything else that might explain the interviewees’ choices. The debriefing questions generally followed the schedule (see Appendix 5.16); however, interviewees who made interesting comments in the think-aloud task were probed further to explore a full range of ideas (for example, reference to particular familial experiences). At the end of the interview, interviewees were specifically asked to consider for a few minutes and see if anything else came to mind.

All the questions used were open-ended to encourage the interviewees to talk more descriptively and develop their own narrative. Prompts were used if the interviewee went silent. The debriefing questions were also used to generate more detailed data when the interviewee was less vocal in the think-aloud task. Throughout the data collection process, the topic guide outlined in the interview schedule remained unchanged.

All interviews were digitally recorded and notes were also taken when it was apparent the transcripts would not reveal sufficient information (for example, when interviewees pointed at attributes or alternatives).

### 5.5.2 Data generation

The interviews were transcribed verbatim and supplemented with field notes where appropriate (for example, when an interviewee had pointed to something then notes were used to identify the object – usually a particular level). For some interviews, the transcription was done immediately after the interview (n=5). Others were transcribed by a professional company with experience of the transcription of health-related interviews (n=14). The professionally transcribed interviews were then double-checked by CV for accuracy by listening to the recording and reading the interview reports, and, where appropriate, supplemented with field notes from the day.

Data were transcribed verbatim in order to maintain close contact with the raw data. When looking at cognitive processes and comprehension, any discourse, even pauses or
mumbles, were sources of useful qualitative data, adding realism to the quotes provided in support of the identified themes.

NVivo® qualitative software for research (QSR, 2012) was used to import the transcripts and also store the audio files in preparation for the analysis.

5.5.3 Process of analysis
The first stage of the analysis involved listening to and re-reading all the interviews and transcripts to ensure complete immersion in the qualitative data. This process cannot be explicitly reported here as it was a psychologically-internal procedure. However, familiarisation with the data simply by reading and re-reading transcripts and listening to audio files kick-started the initial analytical process. During each reading, notes were made and, as interviews were being conducted concurrently to the initial stages of the analysis, points of particular interest were also jotted down in subsequent interviews.

The data generated from the think-aloud exercise were pooled with the data from the debriefing questions and all were analysed using the same methods. This was largely because there was no pattern to the reporting of information in terms of which method (think-aloud or de-brief question) generated the data (for example, some women revealed in the think-aloud that they found the percentages confusing whereas others only mentioned this towards the end of the interview).

From these transcripts and memo notes, the coding and generation of themes began under a loose initial framework (Gale et al., 2013). A framework can be used to identify important areas a priori in order to facilitate the interpretation of qualitative data by providing a preliminary structure to the analysis (Smith & Firth, 2011). The initial codes were highlighted on the transcript text and involved a constrained form of open-coding under the framework. Although constrained open-coding may seem like a contradiction, the approach was ‘open’ in that all new themes were allowed to develop but ‘constrained’ in that they were restricted to the two distinct, but not mutually exclusive, key concepts: 1) decision strategies; and 2) risk. This approach was taken to focus the analysis and remove qualitative data that was not of interest (for example, comments on the warm-up exercise) but still allowed related themes to develop from the codes. Decision strategies related to any accounts of behaviour or heuristics adapted when completing the choice task. The topic of risk referred to any data relating to the probability attributes, whether this be their thoughts about likelihood, perceptions of risk based on experience, or visualisation of the numbers. The process was driven by the interview data and each transcript was systematically coded, with these codes later developed into themes.
The process of coding was both inductive and deductive. It involved looking at the transcripts inductively under the two broad themes of decision strategies and risk, but was deductive in that it was open to new themes which did not ‘fit’ and were then initially coded under new headings and incorporated into a new framework.

Within the two concepts of decision strategies and risk, coding involved an iterative process of developing new themes and ideas as the process advanced. It was then possible to stand-back and look at the snippets of interviews which seemed broadly related and thus re-define the themes in line with the original framework. This was unrestrictive: the same text could be coded multiple times and generate different codes which could possibly overlap. This stage was key in acquiring a familiarity with all interview data collected during the think-aloud and de-briefing questions. The analytical cycles of suggesting, checking and re-checking to expand and then contract themes continued until no new items or ideas were emerging.

The next step involved developing a coding tree stemming from the original framework (see section 5.5.2). The coding tree helped enable discussions with the supervisory team and facilitate feedback on the interview process, even when the interviews were ongoing. The initial broad range of themes were collapsed and re-organised by merging similar codes, moving or attaching others and creating sub-codes when necessary. In the initial stages, the tree was reflected a web of loosely related codes, with many branches of similar topics and tenuous links. After more meetings, interviews and further analysis, the codes were arranged into a series of more structured themes.

5.5.4 Qualitative boundaries and analytical limitations

A common criticism of qualitative research is the impact of the presence of an interviewer both in the collection of data and in its interpretation and analysis (Shenton, 2004). As described previously in section 5.5.1.2, steps were taken to reduce this source of bias in the recruitment and interview stages. It was suggested to women responding to the advertisement that the interviews could be held in their own home, and participants were explicitly told that they should not worry about talking to the interviewer and to only say what they were thinking.

Pilot work for the DCE (described in section 3.5.8.3 of Chapter Three) generated free-text comments from a small internet panel which suggested some DCE respondents were concerned about the confidentiality of the data (whether this would be used for a private company). Therefore efforts were made to reassure participants that anything they said and the choices they made would be completely anonymised in this PhD and any subsequent
publications. Again, interviewees were given an opportunity to read the information sheet and decline consent before the interview even began.

In the interviews, attempts were taken to remain as impartial as possible and minimise researcher bias. Although prompts were made, if interviewees looked for direction they were reassured that there were no right or wrong answers. With concurrent thought-verbalisation other criticisms of bias from interviews, stemming around recall and politeness, were minimised as interviewees were encouraged to speak their mind at all times. After 12 interviews, two meetings were held with the supervisory team to discuss the transcripts and acquire feedback on the preliminary analysis. Therefore the results presented in the next section are as honest, impartial and objective as is reasonably possible given the nature of qualitative research.

5.5.5 Summary of the analysis and methods

The overall process from sampling to conclusion is described visually in Figure 5.2. The process was fluid and involved assessments, reassessments, supervisory feedback and interviewee input before any conclusions could be substantiated. The content of the ovals represent key stages in the empirical study; the arrows show the processes leading to each stage; and the squares describe the methods used in each process.
5.6 Results

A total of 76 women replied to the advertisements and 23 interviews were arranged. Two women cancelled and two failed to show-up. All women who replied to the advertisements and were sent the information sheets initially agreed to participate. However, some (n=2) had family commitments or became too busy to commit to the interview, and some (n=6) were ineligible to participate in this study because they made contact after all interviews were completed (these women were informed of subsequent research and many chose to participate in the study presented in Chapter Six). Appendix 5.17 describes the recruitment process for this chapter, Chapter Five, and Chapter Six, and the difficulties in attributing a response rate for either study.
5.6.1 Interviewees

The final sample comprised nineteen women. The sample size was guided by this previous research, but, as described in section 5.5.1.1, sampling would have continued if new themes emerged from the data; however, a point of saturation appeared to have been met. Although new participants could have been recruited (and women were actively responding to advertisements), the data generated by the accounts of the last few interviewees replicated that of previous participants so further work would have been redundant. This point of data saturation is not only common, it is also considered to be best practice in qualitative research (Silverman, 2013).

The 19 women were randomised to complete one of the two risk communication formats, with nine women receiving the IAP version and 10 women receiving the PO version. The randomisation occurred via an inbuilt function of the Sawtooth software (in an identical procedure to Chapter Four), therefore the process should not have affected the emerging themes or the qualitative data collected. A summary of the characteristics of each interviewee is described in Table 5.2.
Table 5.2: Interviewee characteristics

<table>
<thead>
<tr>
<th>Identity (ID)</th>
<th>Risk format</th>
<th>Age band</th>
<th>Occupation</th>
<th>Date and Location of Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female 1</td>
<td>%</td>
<td>50+</td>
<td>Analyst</td>
<td>Home 18/09/2014</td>
</tr>
<tr>
<td>Female 2</td>
<td>%</td>
<td>35 to 44</td>
<td>Lecturer</td>
<td>University 15/09/2014</td>
</tr>
<tr>
<td>Female 3</td>
<td>%</td>
<td>18 to 24</td>
<td>Social Worker</td>
<td>University 15/09/2014</td>
</tr>
<tr>
<td>Female 4</td>
<td>†</td>
<td>25 to 34</td>
<td>Teacher</td>
<td>Home 19/09/2014</td>
</tr>
<tr>
<td>Female 5</td>
<td>†</td>
<td>18 to 24</td>
<td>Unemployed</td>
<td>Home 13/09/2014</td>
</tr>
<tr>
<td>Female 6</td>
<td>†</td>
<td>25 to 34</td>
<td>Researcher</td>
<td>University 17/09/2014</td>
</tr>
<tr>
<td>Female 7</td>
<td>†</td>
<td>25 to 34</td>
<td>PhD Student</td>
<td>University 18/09/2014</td>
</tr>
<tr>
<td>Female 8</td>
<td>†</td>
<td>25 to 34</td>
<td>Researcher</td>
<td>University 24/09/2014</td>
</tr>
<tr>
<td>Female 9</td>
<td>†</td>
<td>25 to 34</td>
<td>Unemployed</td>
<td>University 17/09/2014</td>
</tr>
<tr>
<td>Female 10</td>
<td>%</td>
<td>25 to 34</td>
<td>Administrator</td>
<td>University 19/09/2014</td>
</tr>
<tr>
<td>Female 11</td>
<td>%</td>
<td>18 to 24</td>
<td>Undergraduate Student</td>
<td>University 19/09/2014</td>
</tr>
<tr>
<td>Female 12</td>
<td>%</td>
<td>50+</td>
<td>Information technology (IT)/Administrator</td>
<td>University 06/10/2014</td>
</tr>
<tr>
<td>Female 13</td>
<td>%</td>
<td>50+</td>
<td>Administrator</td>
<td>University 29/09/2014</td>
</tr>
<tr>
<td>Female 14</td>
<td>%</td>
<td>50+</td>
<td>Administrator</td>
<td>University 30/09/2014</td>
</tr>
<tr>
<td>Female 15</td>
<td>%</td>
<td>25 to 34</td>
<td>Bar worker/ part-time student</td>
<td>University 30/09/2014</td>
</tr>
<tr>
<td>Female 16</td>
<td>%</td>
<td>45 to 49</td>
<td>Finance officer</td>
<td>University 06/10/2014</td>
</tr>
<tr>
<td>Female 17</td>
<td>†</td>
<td>25 to 34</td>
<td>Administrator</td>
<td>University 06/10/2014</td>
</tr>
<tr>
<td>Female 18</td>
<td>†</td>
<td>25 to 34</td>
<td>Engineer</td>
<td>University 13/10/2014</td>
</tr>
<tr>
<td>Female 19</td>
<td>†</td>
<td>50+</td>
<td>Retired</td>
<td>Home 15/10/2014</td>
</tr>
</tbody>
</table>

On average, the interviews lasted approximately 40 minutes (mean time of recording 37 minutes 50 seconds) with the shortest lasting 24 minutes and 54 seconds and the longest taking 57 minutes and 26 seconds. The interviews took place in September and October of 2014 and most (n=15) occurred in pre-booked meeting rooms on The University of Manchester’s Main Campus. Interviewees could choose the location, but most decided to come to the university. The sample included a range of ages and professions as described in Table 5.2.
5.6.2 Coding tree

The initial coding tree (shown in Figure 5.3) included all themes and sub-themes identified in the initial stages of the analysis. This tree describes the “whole picture” and the full range of views expressed in the interviews before any substantive attempt at reduction. Themes were only loosely grouped together with subtle connections explaining potential links. The original framework is shown by black square boxes, and recurring themes are displayed according to their focus: circles representing feelings and emotions; capsules showing logic and reasoning; the diamonds demonstrating expressions relating to the interviewees’ perceptions or experiences; and the white squares encapsulating opinions and beliefs. The arrows are suggested links and possible drivers of each phenomenon.
Figure 5.3: Intermediate coding tree demonstrating identified codes and themes
The initial coding tree was then simplified by reducing the themes into broadly related topics. This involved a back and forth approach of rearranging and establishing connections or patterns in behaviour to create a more coherent account of what occurred in the collected data. The themes are shown in the circles of Figure 5.4 and patterns in behaviour that linked themes are shown by arrows. Figure 5.4 also serves as the structure for the next section of this chapter which discusses the each of the final themes in more detail with supporting, and disconfirming, verbatim quotes as evidence.

**Figure 5.4: Final coding tree identifying key themes and inter-theme relationships**
5.6.3 Experience of completing a DCE

In this next section, the themes summarised in Figure 5.4 are discussed in more detail. The section begins with an exploration with interviewees’ general experience of completing a DCE survey. Section 5.6.4 examines how interviewees used decision strategies to make their choices. In section 5.6.5, interviewees’ reaction to the risk attributes are explored in more detail with comparisons of the differences in the interviewees’ accounts from the two risk communication formats: PO and IAP.

5.6.3.1 Experience of making choices

This section explores interviewees’ general experience of completing a DCE task. Alongside each interviewee quote is an indicator of the risk communication format the woman received and the age bracket they fell into. The ‘%’ identifies women who completed the DCE with risk communicated as a percentage only; and the ‘ʃ’ identifies women who also received visual communication of risk in the form of an icon array.

The interviewees were asked to start to think-aloud as soon as they started the DCE. Many of the initial comments were expressions of surprise or confusion. Interviewees, in both risk communication formats, reported that they had never done anything like the task before and, despite confirming everything was clear prior to commencing, a few asked for clarification of exactly what was expected:

CV: Did you find that difficult?

Participant: Yes, I did at first, you know, before I, sort of, realised what was going on there......by the time I got to the end of it it was like you were almost looking at the same thing on each page. (Female 19: 50+ years, ʃ)

Similarly, other interviewees sought confirmation or reassurance that they were doing the right thing with regards to both the choice experiment and the think-aloud task:

Participant: Sorry, is that wrong?

CV: No, it’s not wrong at all; there are no right or wrong answers. (Female 12: 50+, %)

Participant: Am I making sense? I'm worried! ...Hmm I think I will go for A. Am I making sense? (Female 2: 35-44 years, %)
Previous think-aloud studies have focused on respondents’ reactions to the dominance task (Ryan et al., 2009); however in this study, no women were confused by the dominant choice set incorporated as a test of internal validity, and all participants gave the expected response:

**Participant:** So that’s all a win, win thing there. (Female 13: 50+ years, %)

**Participant:** Oh well that’s really easy... that’s pretty obvious. (Female 18: 25-34 years, †)

When interviewees were asked whether they would make the same choices in real life, the majority confirmed that they probably would:

**CV:** Ok, so do you think if you were to make these decisions in real life, your choices would be the same?

**Participant:** I think, I think that I just chose the ones that personally I felt I would go for. (Female 2: 35-44 years, %)

**Participant:** Yep, I’m not 100% but I think I’d be quite confident. I’m trying to imagine, if I got a letter, what would I do? I think this is what I would do. So I’m quite confident. (Female 4: 25-34 years, †)

However, a few interviewees found it hard to get around the hypothetical nature of the DCE:

**Participant:** You see, that’s why it’s difficult, because you’re not actually presented with it now, as in real life kind of thing, it’s difficult. (Female 16: 45-49 years, %)

**Participant:** Oh, ok, I just think until something actually happens.. I mean even with role play.. you don’t actually know until it happens.. you’ll have lots of other things to think about... because it’s real and it's happening...you might think.. oh..oh well.... oh that's why I hate role play. (Female 3: 18-24 years, %)

The above quotes highlight that a DCE is not an easy thing for people to complete and that it takes respondents a while to familiarise themselves with the survey despite reading the pre-information and having an interviewer present to answer any questions. This DCE was ‘simple’ in that there were only three attributes, but women still expressed a need to take
time to comprehend exactly what was asked, understand the good described and imagine the reality of their hypothetical choices.

5.6.3.2 Experience of completing a think-aloud task

Some women appeared to find it difficult to think-aloud. It was not possible to distinguish whether this was down to confusion with the task, because they could not verbalise their responses, or because they genuinely felt as if they had nothing to say. As detailed in the methods section of this chapter (section 5.5.1.2), prompts were used extensively in some cases but, even then, some women struggled to verbalise:

CV: What about in this case?
Participant: [Pause]. Same again. I’d go for programme B.
CV: Okay. Is that the same...?
Participant: For the same reasons. (Female 12: 50+ years, %)

Women who were hesitant or quiet in the think-aloud task, often revealed more about their choices and strategy in the debriefing interview. One interviewee waited until the very end of the interview to explain more about what was influencing her choices:

Participant: Now we’ve finished, I should tell you that I recently started HRT [hormone replacement therapy] and I know that increases your risk of breast cancer so that probably made me always choose screening. It’s a risk, and I know that, but at the end of the day I’m prepared to take it because it’s made me feel so much better. (Female 1: 50+ years, %)

5.6.4 Evidence of economic phenomena

In their responses, many women expressed behaviours that have resonance with established economic phenomena. The economic behaviours were not reported in the qualitative accounts of the studies identified in the systematic review of DCEs presented in section 5.2 and Appendix 5.1, and so were unexpected and entirely driven by the qualitative data collected. The women did not talk in economic terms, but their self-reported choice strategies could be aligned with recognised economic theories. Whilst these behaviours have not been extensively explored within a health economics context, or in a DCE framework, they have been identified by environmental economists and in studies using the contingent valuation method (Hanley et al., 2003).
5.6.4.1 Opportunity costs

Opportunity cost is a key concept underpinning health economics’ overall aim of ensuring the efficient allocation of resources. This concept is also key to the supporting theories of DCEs which require respondents to trade-off in order to estimate their underlying preferences. Although not explicitly stated in academic-economic terms, many women expressed that they were making choices consistent with the idea of opportunity cost in terms of financial analogies:

**Participant:** You don’t miss, you know, £5 here, £5 every time you go for breast screening, or the ticket cost on the bus or car parking. You don’t miss that, because you’re having a service done... you’re asking as you pay, like you would to go shopping or to go to town.... Bus fare, car, you know, car parking, it’s nothing. You wouldn’t think twice of doing that to go shopping, so why not do it to have your life saved. *(Female 12: 50+ years, %)*

**Participant:** I think it’s very important and spend more money on healthcare than other things in life probably, but that’s just probably me personally. *(Female 11: 18-24 years, %)*

Another emergent theme was that women seemed to be thinking in terms of ‘efficiency’ of, or the opportunity costs to, the NHS more generally. In a publically funded health care system, many women seemed to be concerned about a programme being wasteful:

**Participant:** Well it seems to me like programme B sounds like it’s much more efficient in what it’s doing, if there’s no women detected who had unnecessary follow up. *(Female 14: 50+ years, %)*

**Participant:** See this is a waste of a lot of resources. I think I would go for none of these... This is too expensive and the other is too wasteful. 20% of people coming back is too high, it's not so good... I’m evaluating the programmes so first I see which would I go for then my second point of view is if it’s useful or not useful. If programme B is calling half the people back with follow up then it means there's a lot of wastage on the system. *(Female 2: 35-44 years, %)*

Two women also expressed ideas about ‘regret’ contributing to their decision-making:
**Participant:** It doesn’t really matter about the cost because obviously if you have breast cancer and you haven’t gone to Programme A or B then you’d be kicking yourself, you know, if you were to end up terminally ill... *(Female 11: 18-24 years, %)*

**Participant:** I think there’s quite a few things, but for me I think a big concern would be being treated unnecessarily because you can’t really reverse it once you’ve started the treatment. *(Female 6: 25-34 years, ♂)*

### 5.6.4.2 Discounting

It is well established in neoclassical economics that people have a preference for money today rather than in the future and thus discount returns at a later-date (Frederick et al., 2002). It is less clear how people treat non-financial benefits such as gains to health (Torgerson & Raftery, 1999). In this study, there was evidence that some younger women were discounting the benefits of breast screening:

**Participant:** But also I think - I know I'm only sort of, like 19 years away from starting such a programme, but it seems like a long way away. Dying doesn't seem to me - that's not something that's going to happen to me anytime soon. ... *(Female 18: 25-34 years, ♂)*

One contributing factor to a discount rate appeared to be the uncertainty surrounding both their future income and future health (ultimately uncertainty in their future budget constraints and the future utility they could acquire from participating in the screening programme):

**Participant:** Obviously you don’t know if you’re going to have £1,000 just to spend over 20 years. *(Female 11: 18-24 years, %)*

**Participant:** You know, where I’ve been in and out of hospital, had surgeries, and different things, lots of different things. So, you never know how you’re going to be at any age really, do you? *(Female 19: 50+ years, ♂)*

When consulting breast cancer experts in the original design of the DCE, they expressed an interest in understanding the drivers behind women’s decision to attend screening. This was deemed to be useful in terms of predicting adherence and explaining why some women coming to the end of the programme chose not to attend (Edwards & Jones, 2000).
Some women’s statements indicated that their choice was probably only reflecting their preferences for early screens, and as they reached the end of the programme they may no longer choose to participate. This is not discounting, but rather women updating their priors about the expected utility from screening (the probability of a positive test result) given their previous results:

Participant: I think it would also depend on if I’ve had one or two before and they’ve turned out okay, I think I would be willing to pay less for further ones but I would be willing to pay probably the most for the first few. So that’s how it would be... I think my views might change if I’ve had three and it turns out they’re all alright and I’m almost...yes, I’d be almost 60 (Female 6: 25-34 years, %)

Participant: I would always go, but dependent on what it would say would influence my decision to go back. (Female 4: 25-34 years, %)

5.6.4.3 Decision process and model selection

The systematic review of qualitative research methods (reported in section 5.2 and Appendix 5.1) found that some authors had used qualitative data to inform the quantitative analysis of choice data (see Pitchforth et al., 2008). Although this is not touched upon in any of the DCE guidelines (Bridges et al., 2011b; Lancsar & Louviere, 2008), some respondents revealed behaviour that could inform model selection through their treatment of the ‘No Screening’ alternative. Some interviewees reported that they first considered whether to opt-out of screening before even considering the two screening alternatives:

Participant: So no screening, yes, okay. I think Programme B is probably preferable to that one, yes. (Female 10: 25-34 years, %)

Participant: For this one I’m not really sure because I think that like Programme A the smallest one has ten per cent who are treated unnecessarily compared to the twenty but then now it’s a case of whether it’s better to not have a screening at all than to be one of those ten... (Female 6: 25-34 years, %)

5.6.4.4 Information attendance

In terms of attendance to the information presented in each choice set, it was clear that some interviewees were not taking everything into account. One woman expressed
complete surprise when prompted about her thoughts of the ‘No Screening’ alternative having previously ignored it:

**CV:** Ok, do you consider this [opt-out] when you make your choices?

**Participant:** I didn't even see this! So what.. so what is this? No cancers detected.. Wait...so this... what? I don't get this. So this...?.. [paused] so you have no cancers detected...because you wouldn't be screened?

**CV:** Yes...

**Participant:** That none-option doesn't even come into it. (Female 3: 18-24 years, %)

For another woman, the opt-out offered a way-out of making a choice at a point of indifference:

**Participant:** I think I’m probably in the not sure and I’d probably in this one go no screening for this one. (Female 6: 25-34 years, ♂)

Many women expressed that they would always opt-in to a breast screening programme. This was not surprising given the large and positive ASCs identified in five out of six of the preference-classes described in Chapter Four. The quotes below further support this quantitative finding:

**Participant:** That wouldn’t be an option for me at all...Yes, sorry, I hadn’t even bothered with that one because you know, I would definitely go, if it was me I would definitely go for screening, no matter what, even if they didn’t...it was unnecessary and I had chemotherapy and whatever, I would still have gone for it anyway. (Female 14: 50+ years, %)

**Participant:** So there’s no screenings at all? No I’d never go for that....So this third option is always going to be with no screening?

**CV:** Yeah.

**Participant:** Well I’m never going to go with that then, I can’t do that. (Female 13: 50+ years, %)

Non-attendance was also present in the form of ignoring attributes. This specific phenomenon of attribute non-attendance is explored in more detail in Section 5.6.4.8.
5.6.4.5 Information asymmetry

It is well established that information asymmetry is common in healthcare markets (Bloom et al., 2008). Whilst DCEs are often cited as a way round this, trying to specifically understand patients’ or the public’s demand for a healthcare interventions, many women expressed that their choices were completely dependent on the advice they received from government agencies or ‘experts’:

**Participant:** So I would be guided by whether I thought there was a reason. If I was invited, I would definitely go... But again, I'm thinking if it was offered to me then there's a reason it would be offered to me. I would, I think I would want to go. If the NHS said oh, we think you should go, then I would be like ok... I would trust them. Only because I don't know anything about it myself, if I had a better judgement then I would make a choice for myself. But I don't know much so I would be guided by them. *(Female 4: 25-34 years, )*

**Participant:** Obviously the NHS has decided for a reason that 50-70 is a good age to go so I think I would go for it... One thing I find is, you don't often receive information on how accurate any screening is or how effective the medication is, so it's difficult for people to think - like when the doctor tells them something you don't really know is this going to work? Is this giving me the result I want? Because you're putting so much trust in that and most people don't have the time to go off and research it themselves. *(Female 18: 25-34 years, )*

**Participant:** You don’t think about it being necessary or not. And you just, I suppose, with breast, you just kind of do what your GP tells you, don’t you, a lot of the time. And you might have limited information about what you’re supposed to do, and it’s just kind of, well, I should go for that, so I should go for it. *(Female 16: 45-49 years, %)*

**Participant:** Yeah, I think basically because it’s just the normal thing to do as well, people do it and it’s just like... *(Female 15: 25-34 years, %)*

Whilst many women expressed feelings of trust, some women were surprised at the statistics presented and felt that they had not acquired sufficient information about screening:

**Participant:** I thought the percentages of detection were quite low, yeah, it just sounds quite a low number isn’t it to go through all that and just, it’s quite...so I’m quite
surprised about that, I thought...Yeah, but yeah, I’m surprised the detection levels aren’t higher and, yeah...mm. Yeah. I’ve learned stuff that I didn’t know about. I didn’t realise that they took a...up about the unnecessary follow-ups actually. (Female 15: 25-34 years, %)

5.6.4.6 Familiar good effects

In the ecological literature, there is empirical evidence that valuation of animal species can be influenced by people’s prior attitudes towards the wildlife or natural resource and its familiarity (Czech et al., 1998). Potentially, higher valuations could be a consequence of social, political and media attention to the subject (Czech et al. 1998). Whilst this has been briefly discussed in some health economics articles (where the success of breast cancer charities in their campaigns and subsequent influence on policy has been noted (Pennery, 2007)) it has not been specifically investigated in a health DCE context. However, cancer, and breast cancer in particular, appears to have some socially-constructed virtue ascribed to it by some of the respondents in this study. Whilst these media effects and charitable campaigns are established in terms of revealed actions (Evans et al. (2014) highlight the ‘Angelina Jolie effect’ on breast cancer referrals) it was unclear whether this influenced hypothetical choices. At the time of the interviews, a celebrity was diagnosed with breast cancer and a few women drew upon the celebrities and/or the media in their debriefing questions:

Participant: Like Angelina Jolie, isn’t it? If you compare with that. (Female 13: 50+ years, %)

Participant: Michaela Strachan... She’s just had them both off, hasn’t she, because of breast cancer. She’s not that old, is she? (Female 12: 50+ years, %)

5.6.4.7 Healthcare perspectives and non-use values

As introduced in section 2.2 of Chapter Two, a key stage in understanding how people value health care is to understand the perspective they take and whether the valuations derived are welfarist or extra-welfarist (Brouwer et al., 2008). In this DCE, women expressed sentiments implying they were completing the DCE from a population-health (rather than individualist) perspective, thinking of the greater benefit to the general public:
Participant: the higher, the more people we can, you can find, the more people you could treat, the higher the chances of survival, so... I think that on a humanitarian level, that you’ve got to open it to everybody, and try and save as many people as possible. (Female 13: 50+ years, %)

Participant: if it’s like a mum with children, if you’ve got kids or you’ve got a big family then you’ve got to look after your health and you’ve got to try and get medical advice and see if there’s any problems. Or if they couldn’t... some people have it really bad have got families, one parent families, having to do loads of stuff in debt and it’s just like they literally can’t afford to take a day off work. So that’s...yeah, that would be a bit of a factor with people that actually come and get these tests I think. (Female 15: 25-34 years, %)

There was also evidence to suggest that the respondents were not just considering other people’s health whilst they were completing the DCE, but they were specifically drawing their own utility from these altruistic provisions. In the environmental economics literature (Hanley et al., 2001; Perman et al., 2003), these philanthropic motives are identified as potential ‘existence values’ which contribute to the non-use values people may have from a good or service:

Participant: So it’s difficult to say, because although you don’t want lots of people having unnecessary treatment, at least if it’s more successful in detecting it, I’d probably for A there then. (Female 16: 45-49 years, %)

Participant: It’s tricky because until you’re in that position I don’t ever want to have to put 20 per cent of people through any unnecessary follow ups.. hopefully not submitting people to, as many people as possible obviously, hopefully not submitting the highest number to any unnecessary procedures but outweighing that and balancing that off against what would be mostly detected. It’s not the perfect solution, that’s the frustrating thing, you can never tell... I suppose part of my influence is also the wider community. (Female 10: 25-34 years, %)

There was further evidence of existence values in line with ‘warm-glow’ theories (Kahneman & Knetsch, 1992). Receiving a warm-glow suggests some women gained utility from the moral satisfaction of having an inexpensive service that anyone could use:
**Participant:** Would you get more people going in for Programme A despite the fact there’s 20 per cent unnecessary follow up. There is a higher rate of detection but is that balanced off the fact that it’s only £20 per screen. We were estimating if 100 people going to each of these programmes... if we were to look at the general population as a whole and offer the population these two programmes only a certain number of people would be able to afford Programme B...because Programme A’s cheaper you’ve got more people going for screening on Programme A and therefore you’re finding a higher rate of cancers detected in any case... At a cost of £20 it would hopefully be far more of an incentive to go along and spend that money to have the screen...Thinking about...certainly people that I work for whether they’d have the disposable income to do it as well is slim to none. Therefore the lower the cost the more likely it is that they would go. Certainly if you’re talking about hundreds of pounds per screen I think that would be far outside the reach of most people. (Female 10: 25-34 years, %)

**Participant:** Even though I personally would still pay £200 to get the same thing, I think that for the population, and considering the demographic of the people who get the screening, that £200 might be too much for people of that age maybe. So to encourage more people to go for it I would go for programme B in that case. (Female 8: 25-34 years, ♂)

Non-market valuation methods could be susceptible to bias from respondents gaming the survey to give a desired answer, or refusing to complete the survey because they disagree with the principal entirely (Smith, 2000). In environmental economics, one potential reason for these strategic responses is that respondents have a strong belief that something should be preserved (Perman et al., 2003). In health, the NHS and the NHSBSP are the items to be protected. Whilst there were no explicit statements expressed by any of the respondents in this study, some women did indicate protective views of the NHS:

**Participant:** I mean, I honestly believe in the NHS. They’ve done so many good things for me in the last 20 years and I would prefer to support the NHS every time and, obviously, there’s a few cases where you’ve got to jump the queue because you’re desperate...and I think the NHS are suffering at the moment. (Female 19: 50+ years, ♂)

Strong beliefs about protection could lead to another bias: protest responses. A protest response reflects a choices given on principal and therefore may not reflect the
respondent’s true value (Jorgensen et al., 1999). One statement indicated a temptation to
give a protest response:

**Participant:** I’m also trying to think about, for example, my gran having to go for one
and having to pay £200 each time. She probably would, but I just think that it’s quite
mean. She shouldn’t have to do that, kind of thing. (Female 8: 25-34 years, ♂)

### 5.6.4.8 Non-compensatory preferences

A number of key axioms support this rational choice model and estimation of utility
requires respondents to express reflexive, complete, transitive and continuous preferences
(Varian, 1992). If a respondent exhibits non-compensatory preferences and no amount of
money can compensate them for not having screening, they will choose to go for screening
at any cost. When respondents are completely unwilling to trade-off this cost attribute they
are exhibiting attribute-non-attendance. This violation of the continuity axiom through a
reluctance to trade-off the cost attribute was common (particularly in women aged 50+):

**Participant:** I still think that having the screening totally outweighs what the cost would
be anyway….well because I don’t think you can really put a cost on your health and
your life, your survival. (Female 14:50+ years, %)

**Participant:** So I just can’t see how you can go wrong by spending money on something
because money against life [laughter] it’s no comparison… There’s no comparison…it’s
not a matter of money. And I do stress that’s not because I’m well off … generally
speaking, to be told with great relief that you haven’t got cancer, you can’t buy that.
(Female 13: 50+, %)

**Participant:** Well, where it’s saying cost to you I don’t think the cost should come into
it…it’s your life you’re talking about. (Female 19: 50+ years, ♂)

**Participant:** I just don’t think you can put a price on your health. (Female 8: 25-34
years, ♂)

**Participant:** It’s a lot of money but then I think it’s hard to put a price on cancer being
detected, isn’t it?… just think that with stuff like that, money’s not a factor..you’ve got to
pay for screening then I’d pay it. (Female 17: 25-34 years, ♂)
Even when women took into account the cost attribute, many would exhibit lexicographic preferences by simplify the task by rank-ordering the attributes. The most common rank-ordering involved making decisions with the preference: probability of detecting a cancer $\gg$ risk of unnecessary follow-up $\gg$ cost of screening:

**CV:** Did you find yourself sort of concentrating on one of the characteristics or...?

**Participant:** Yes, the percentage of detected more than anything else, really....And then the unnecessary follow-ups and then the money was my third kind of thing, but yes, it was governed by percentage of detected from the screening. (Female 17: 25-34 years, †)

**CV:** So, you’re just were you going for the thing with the highest detection rate, best at detecting?

**Participant:** Absolutely, yes. (Female 19: 50+ years, †)

Whereas other respondents were able to interpret the cost attribute and made choices in line with the underlying economic theories:

**Participant:** It sounds silly because it’s obviously your health and you can’t put a figure or price on it but people will do because that’s the day to day life that we lead. If you’ve got to go home, feed the kids, clothe the kids or whatever and if they haven’t got the money to do that then they’re not thinking much beyond that I don’t think. (Female 10: 25-34 years, %)

However, worryingly, a few women also associated the cost of the screening programme with its quality:

**Participant:** Oh I think there’s an element of saying, if it’s £20 a screen then I’m bound to think that’s not serious: it’s a no-good screening programme that’s basically just, you know, making random decisions and I don’t think it's working. (Female 2: 35-44 years, %)

**Participant:** Even the unnecessary treatment I’d rather be on the safe side than think I’ve gone and I’ve got something and it’s not been detected because it’s cheaper or because of another reason. (Female 11: 18-24 years, %)

**Participant:** It’s like why wouldn't I maybe pay a bit more to be seen, sort of, in a more beneficial and professional way perhaps? (Female 18: 25-34 years, †)
One woman also drew upon her experience of paying for health care to add realism to the hypothetical choice and estimate how much she would really be willing to expend:

**Participant**: Funnily enough I was looking into like the HPV [human papilloma virus] vaccine but they said if your cervical smear test didn't come back abnormal they wouldn't test you for HPV, so you could go home and buy the vaccine which was £400... but they couldn't tell you whether or not it was going to work ... and I thought, mm, £400 for something that I've no idea whether it's going to work or not is just too much money for me. But if it was I'd say up to about £80 I probably would have done it anyway, so it's kind of like similar... *(Female 7: 25-34 years, f)*

5.6.4.9 Stability of preferences and anchoring

In some stated preference studies, the axiom of complete preferences is tested with a stability check (Czajkowski et al., 2014). In this study, some women reported that their preferences changed or became more defined as they completed the experiment (potential learning effects):

**CV**: Do you think your choices changed as you went along through the different choices?

**Participant**: Yes, I think so. I think I started to think more about the cost and whether I would prioritise being treated unnecessarily more than the chance of it being detected and treated. Most of the time I was on the treated unnecessarily, that was the main thing for me, but then when I started to think about the fact that if you have cancer then it might be missed that became more of a priority as well. So, yes, I think it changed a little bit. *(Female 6: 25-34 years, f)*

**Participant**: I think, maybe at the start, I was more focussed on the price, but then actually having gone through a few and weighing it up it's actually more important to look at the other factors rather than the money. *(Female 9: 25-34 years, f)*

Whilst other women felt that they were consistent and did not change their decision rules:

**Participant**: I think I was consistent. I think it got difficult when one was really expensive and one was quite a lot cheaper. Again, for the same reason that I can’t expect everyone to do what I would do, kind of thing. *(Female 8: 25-34 years, f)*
Participant: I think I based it on the same reasoning but in weighing up the options it became easier to understand what the options could be. I think I stuck to the same logic all the way through though. (Female 10: 25-34 years, %)

The DCE training materials were extensively piloted (as described in section 3.5.8 of Chapter Three) and included a description of all the attributes and potential levels prior to starting the DCE. However, some respondents appeared to anchor their choice based on the levels which had appeared in previous choice sets:

Participant: Oh, is that [£1,000] the highest?
CV: Yes
Participant: Oh ok the earlier one where I chose a different one even though it had a high cost. But now this high cost is what's stopped me here.
CV: Yep. Do you think you change your mind as you see new choices?
Participant: Yes! Not massively, but it's making me think more. Like £1,000 is massive! And it makes me want to choose that one ... Weird how it makes you think. (Female 4: 25-34 years, %)

5.6.5 Experience of completing a DCE with risk attributes

A key aim of this PhD is to investigate how DCE respondents’ understood risk attributes. This next section looks at the think-aloud interviewees’ reaction to the risk attributes specifically, before comparing how these differed between the communication formats.

5.6.5.1 Computing Behaviour

For some respondents, the use of risk attributes appeared to induce behaviour indicative of calculating and a desire to ‘work out the answer’, and this was more apparent in women who received the PO version:

Participant: I was trying to do the maths...[paused]
CV: Ok, so when you try and ‘do the math’ what do you mean?
Respondent: Well... I think if you actually worked out the maths, I might be slightly out. (Female 3: 18-24 years, %)

Participant: Oh because I need to compare, I'm making decisions to make them comparable. So this is 7% so I'm halving it to 3.5% which is nearer the 3% and then I'm halving the rest to make it equal and then I can see it's half and half. Then I'll look at the cost - and the 7 - 5 ratio - and then the cost. (Female 2: 35-44 years, %)
Participant: So actually yes, so if you've got - I'm just trying to work this out. You'd have approximately a 3 in 10 chance - no you wouldn't at all, you would have a 1 in 3 chance, sorry, of getting a false positive there and you'd have a 0% chance there - 7% or 10%.

(Female 18: 25-34 years, ♂)

Some of these calculations were more apparent with women receiving the IAP version, suggesting the format might aid trade-off decisions:

Participant: It's like £50 versus £150 and then the follow up looks like about half, there's only five per cent who'll be treated unnecessarily whereas it’s double the amount of Programme B... Then looking at the number of people who have had it but then at the same time the cost is a quarter of what it would be. Programme A is a quarter of the cost of Programme B but because the amount’s so different in terms of the financial cost for this one I would pick A. (Female 6: 25-34 years, ♂)

Participant: I suppose because I work in engineering I do tend to consider things quite carefully and I'm quite logical in how I make choices. As I say, I'm probably more logical than most other people. I'm sure that things like this can probably trip me up too. (Female 18: 25-34 years, ♂)

Two women (who both received the PO version) reported that they were attracted to the largest number first (whether that was probability of detecting a cancer or risk of unnecessary follow-up) and would make a decision from there:

Participant: the higher the number, I was looking for there. (Female 14:50+ years, %)

Participant: Ermmm the high percentage.. and then I think hmm .. then the money... and then that doesn't come into it. (Female 3: 18-24 years, %)

One woman reported flipping the negative risk into a positive one to make the comparison between attributes easier:

Participant: Errr.. actually..well 80% wouldn't have an unnecessary follow-up...so I'm just thinking..[paused]..oh actually... (Female 3: 18-24 years, %)
5.6.5.2 Experience and analogies

Many interviewees drew upon their previous experiences to determine their own perceived risk. Women who had no history seemed to believe they were at low risk:

Participant: I suppose, like, cancer is not in my family, either breast cancer or any kind of cancer is not in my family at all, so it’s something that I kind of don’t think about (Female 17: 25-34 years, ♀)

Participant: No-one as far as I’m aware has any history of breast cancer in my family. So I don’t know if what I would do is different to someone who does. (Female 6: 25-34 years, ♀)

But women who had a family history also reported that they did not feel at increased risk:

Participant: I mean my grandmother died quite young of breast cancer, but I still think my risk is probably average because I think that it’s quite a common thing to happen. (Female 18: 25-34 years, ♀)

Participant: I don’t think it influenced my choices. And at the moment I really don’t think I’m at risk at all. (Female 4: 25-34 years, ♀)

The quotes above illustrate a phenomenon known as ‘optimism bias’, where prior information does not affect perceptions of vulnerability to a particular risk (Weinstein, 1984). A noticeable exemption was when one woman described how her mum’s experience with breast cancer did influence her choices:

Participant: My mum went to a van...can I talk to you about that?
CV: Yeah.

Participant: My mum went to a van, just from work, they had that, and they detected her cancer And three years later the van came round again, and she went. And it was a young girl who’d never...like she was new to it. And she said to my mum, I don’t know whether I should do you because you’ve no breast there. So she said, well, rather than waste your time coming back, I’ll do what little bit you’ve got and then you know I’ve done it...saved her life that, because apparently they don’t do it...and she’d got cancer again in it. And they said, you owe...this is at the Nightingale Centre, and they said, you owe this girl your life. And they caught it, and she had to have it completely removed, lymph glands. So it was worth everything just for that.
CV: Do you think that influences your choices?

Participant: Yeah, a lot. Because she could be dead but for that young girl who didn’t know. (Female 12: 50+ years, %)

However, the experience of close friends appeared to have influenced interviewees’ preferences for screening:

Participant: A lady I know, she had a mammogram earlier than 50 ... and they detected breast cancer. ...And she has been really ill with it and lost her hair and everything else. And she got very depressed and everything. She was a very strong woman. So I think when you see somebody like that, it does worry you a lot more, thinking, because it does bring it home to you that it could happen to you. (Female 14:50+ years, %)

Other women felt that risk was just ‘a fact of life’ and which could happen to anyone:

Participant: Life is a risk. Everything we do is a risk. (Female 13: 50+ years, %)

Participant: So sometimes there's good, and sometimes there's bad. So I think it's kinda like, it goes along with the screening, you know. (Female 2: 35-44 years, %)

In addition to breast cancer, one interviewee looked at her experience of other ill-health to think about her preferences for the benefits and risks of screening:

Participant: Well, I don’t know. I’ll tell you something I have actually got on my mind. My boyfriend’s mum and sister have got Coeliac disease, so gluten intolerance, and he thinks he’s got it, and he won’t go to the doctors... like you can’t put a price on your health, and I know it would be rubbish to have Coeliac but you’ll feel so much better knowing and you can deal with it then. So I guess that is just another example of where I think it’s better to know so you can deal with it... (Female 8: 25-34 years, %)

Some women also drew upon their experience of previous breast screening or other screening programmes in order to make their choices:

Participant: I used to have a little bit of a phobia about going for a smear test... It's like if I can avoid it with good reason and it seems like a good reason to - it's like I can justify it. Well this is unreliable or I would worry then that - although it's a bit of a crappy reason, it's still a justification and it's like I feel like I could back that up in my own mind. (Female 18: 25-34 years, %)
**Participant:** I suppose, I don’t really know why it is but I, sort of, like - it’s like coming back to the smear test thing. It’s like I don’t really like seeing doctors very much. It’s like if I can avoid it with good reason and it seems like a good reason to - it’s like I can justify it. Well this is unreliable or I would worry then that - although it’s a bit of a crappy reason, it’s still a justification and it’s like I feel like I could back that up in my own mind. *(Female 4: 25-34 years, ♀)*

To aid their interpretation of risk, some women used analogies to compare the magnitude of the different levels. Interestingly, one woman decided 20% was sizeable as she would be happy with a 20% discount in a shop:

**Participant:** It’s royalties on a book. It’s 20 per cent off something as in discounts, it’s a fairly high percentage compared to most discounts that you’re offered. *(Female 13: 50+ years, %)*

For other women, their interpretation of the magnitude of risk was quite difficult to verbalise:

**Participant:** So I like to, sort of, think about let’s just say you’re more likely to be eaten by a shark than struck by lightning or whatever, because it’s fun. I like things like that. **CV:** Did you find you compare it to any sort of likelihood of other events occurring or do you just…

**Participant:** Not really but I did notice that like I say, some of - I mean all of the figures seemed fairly low to me. So 3% seemed extremely insignificant, but I would say is that if it was the chance, I don’t know why I thought like this, but if it was the one where it’s a chance of discovering actual cancer, the lower figures didn’t bother me… I can’t explain that. They just seemed to unnerve me a bit more and I don’t really know why. *(Female 18: 25-34 years, ♀)*

**5.6.5.3 Visualisation of risk**

Women in both the PO and IAP versions of the DCE often read out-loud the risk attributes as natural frequencies (such as ‘1 in X’):

**Participant:** Oh well I’m thinking obviously they’re both 7 per cent but 1 in 10 people will have unnecessary worry in A, whereas 1 in 20 in B *(Female 11: 18-24 years, %)*
Participant: One in a hundred people who will have an unnecessary follow up, ten in a hundred, yes. (Female 10: 25-34 years, %)

However, this was largely done for the risk of unnecessary follow-up which contained easier levels, rather the detection which contained more complex probabilities (particularly 3% and 14%) which were more challenging to convert into a 1 in X number:

Participant: Because it’s like 1 in 5 people are going to end up being told, oh we think there’s something like suspicious or not quite right...this one [pause], oh well I’m thinking obviously they’re both 7 per cent but 1 in 10 people will have unnecessary worry in A, whereas 1 in 20 in B, yet B is like four times more expensive than A. So B due to money. (Female 11: 18-24 years, %)

Participant: I think to avoid like unnecessary stress and unnecessary procedures which was one in five women, would go through, then I would pay extra money for it... I think 20% is really pushing the limits, because that’s one in five people who have to undergo unnecessary surgery. (Female 7: 25-34 years, ♂)

Participant: I think saying that something happens a fifth of the time that’s quite - it's a more meaningful statement to me I would say. (Female 18: 25-34 years, ♂)

Without prompting, many women justified the translation of percentages into frequencies as an easier way of understanding the risk and aiding visualisation:

Participant: I’m just trying to picture it. Sounds silly, but like the people. So if it was 100 people it would be three people who found out they had it and it would be 20 people potentially going for chemotherapy, which would be really stressful for them. (Female 8: 25-34 years, ♂)

Participant: I just imagined literally one in ten. There’s ten and then you’re just like one of them and the odds of it all, one in 20 or one in five that kind of thing. So I literally imagined just the number of people rather than how much it would work out in terms of hundreds and thousands of people. So yes, that’s how I thought about it in my head. (Female 6: 25-34 years, ♂)

Participant: So I could visualise little people! But...like in children’s books or whatever, but numbers I do have a problem with, especially when the numbers, you know, like 14
per cent and ten per cent, doesn't seem that much of a difference between them... Yes I do find that difficult... Even though this looks like simple numbers, but me trying to... because I think in pictures I think. I'm trying to visualise it. And I'm finding that difficult. I don't know, I mean for me I'm trying to look at it as though they were people. Little pictures of people there, that's like if there was ten people and five people... (Female 14:50+ years, %)

For two women this was very personalised, and one of these imagined herself in a group of five close friends:

**Participant:** Yeah, like because I've got a group of five friends, you see, we're in a five, so I always think in that case one of us. That comes to my head quite a bit. (Female 11: 18-24 years, %)

**Participant:** I think because I'm thinking of seeing people, I know personally, who've had breast cancer detected and are undergoing treatment, and to me three per cent sounds like a very low number of people detected with having cancer. (Female 14:50+ years, %)

It was very apparent that for many respondents receiving the PO version, there was an initial alarm over the numbers presented, but some women felt that they were able to overcome this as the task progressed:

**Participant:** I'm thinking, oh God, numbers, percentages! ... I don't find numbers easy to deal with... For me, I mean, I've never actually been that good at maths or anything, and perhaps I've got a bit of a number blindness I don't know, but I feel a bit stressed when I see numbers... I knew what my options were and I was quite clear in what I wanted to choose, but it was the numbers that I found difficult to work with... (Female 14:50+ years, %)

**Participant:** Yeah... but then I was trying to work it out... I was getting confused with the percentages... (Female 3: 18-24 years, %)

**Participant:** It's just personal to me I think, but it was quite a lot of figures and when you sort of look at one and then look at the other, and then it clicked with me and I was fine, but I think I needed perhaps a little bit more time to understand the percentage. (Female 13: 50+ years, %)
One woman who received the IAP explicitly commented on how it aided her visualisation of risk:

Participant: *It's easy to visualise it where you've got the pictograms here because you can quite clearly see it's one line out of five.* (Female 18: 25-34 years, #)

The next section of this chapter will discuss, in more depth, the results presented in sections 5.6.3, 5.6.4 and 5.6.5.

5.7 Discussion
The primary objective of this chapter was to understand more about how individuals complete DCEs which contain a risk attribute and whether this differs between risk communication formats. In section 5.7.1 the main results of the qualitative interviews are discussed in more detail, drawing upon existing literature and exploring the implications of the findings. Ultimately all of the findings rely on the validity of the think-aloud as a method to accurately elicit people’s thoughts. The strengths and limitations of the study, particularly the qualitative approach taken, are presented in section 5.7.2.

5.7.1 Key findings
The key findings from the interviews described in this qualitative study will be discussed in the following sections: in section 5.7.1.1, results relating to the method, think-aloud interviews are discussed; in section 5.7.1.2, the key insights about respondents’ choice strategies are explored; and in section 5.7.1.3, findings related the risk communication formats are presented.

5.7.1.1 Think-aloud Interviews
The final sample size of nineteen women was in line with other think-aloud studies identified in the systematic review presented in section 5.2 and Appendix 5.1 (see Cheraghi-Sohi et al. (2007) who used a sample of twenty; and Ryan et al. (2009) who used a sample of eighteen). Respondents could choose the location, but most decided to come to the university. This could possibly reflect the anonymous nature of the recruitment and the reluctance of interviewees to invite a stranger into their own home.

The think-aloud method proved useful for some women who were able to continuously verbalise their thoughts, however, for others this was more difficult and the debriefing questions were vital in eliciting their thoughts. The think-aloud results provide further
confirmatory evidence that the DCE had validity with many women trading-off the attributes presented and understanding the task at hand, although a few women (in both risk communication formats) reported that it took them a while to familiarise themselves with the task. The unfamiliar nature of the choice task was also reflected in women’s comments on the hypothetical situation and the need for reassurance that they were doing the ‘right’ thing. This serves as evidence of the importance of effective training materials explaining exactly why their choices are of interest and what the task will involve.

5.7.1.2 Choice Strategies

A key assumption in the supporting theories to DCEs is that respondents will behave rationally and choose to maximise their utility. A number of key axioms support this rational choice model and estimation of utility requires respondents to express reflexive, complete, transitive and continuous preferences (Varian, 1992). If a respondent exhibits non-compensatory preferences (no amount of money can compensate them for not having screening), they will choose to go for screening at any cost.

Respondents unwilling to trade-off the cost attribute exhibit attribute-non-attendance. It is satisfactory for cost to have a negligible impact on utility, but statements made by some interviewees implied they were actively ignoring it (even at levels of £1,000). This violation of the continuity axiom through non-attendance to the cost attribute could result in upwardly bias WTP valuations. This is not new; DCE respondents’ reluctance to trade-off attributes in a health setting has been identified in other studies (Lagarde, 2012). Similar behaviour has also been reported in environmental DCEs (see Carlsson et al. (2010)) and studies in health DCEs have found insignificant coefficients on cost (Watson et al., 2009). However, this result was surprising given that the estimated utility parameters for the cost attribute were significant in all preference-classes described in Chapter Four. The discrepancy with the quantitative results could be attributed to researcher-presence and social-desirability bias with respondents suggesting that cost was unimportant because it was what they believed was ‘right’ or in line with ‘social norms’.

A few women who did take into account the cost of screening interpreted it as an indicator of the programme’s quality. This association could also lead to biased valuation estimates as respondents associate higher cost with more utility. The finding could explain why some (see Augustovski et al. 2013) studies have found the cost attribute to have a positive coefficient in their analysis.

Lexicographic preferences were another simplifying heuristic that some women introduced to make their choices. As with attribute non-attendance, the rank ordering of attributes is a
violation of the continuity axiom. Not accounting for lexicographic preferences in the modelling of DCE data may result in bias estimates as the decision strategy introduces systemic errors (Campbell et al., 2006). As this DCE had only three attributes, it is difficult to verify the existence of this simplifying technique.

Another violation of RUT occurred when women made choices using other people’s budget constraints. Concern with the affordability of the programmes to other people was a recurring theme in the qualitative analysis. If DCE respondents employ stricter budget constraints when making their choices, this could result in a conservative estimate of true WTP. This result highlights the difficulty of framing cost in a health DCE. However, the effect could have been exaggerated by the presence of an interviewer and a desire to appear generous or philanthropic.

Linked to feelings of altruism was the existence of non-use values where women expressed that they received utility from the programme that was not related to their direct consumption. For example, women stated they felt it was important to have a screening service that was accessible to others and this affordability (determined by the level of price) also had impact on their choice. This is a key finding as it implies women were making choices to maximise the population’s health, not just their own.

5.7.1.3 Interpretation of risk

Some women expressed feelings of confusion when the respondents were presented with the PO version, and overall the task appeared to be more daunting. One woman even stated: “I’m thinking, oh God, numbers, percentages!” Icon arrays appeared to make the prospect of the DCE task more attractive and engaged women in the survey. Similarly, a few respondents also felt the need to make a calculation and produce a ‘correct’ answer when the DCE was presented with percentages; despite the training materials and interviewer explicitly stating that there were no right answers. This should be accentuated to DCE respondents possibly throughout the task.

Respondents who received the IAP version expressed statements supporting the use of icon arrays with one woman explicitly reporting that their presence improved her ability to trade-off between attributes. As visualisation of risk was common in interviewees receiving the percentages only version, the addition of an icon array appeared to relieve the cognitive burden of ‘imagining women’ in an already imaginary scenario. This is in line with previous studies which have found icon arrays aid people’s processing of screening information (Hess & Siegrist, 2011).
Some women also created analogies to aid their interpretation of the risk attribute (and one woman compared this to a 20% discount in a shop). It is established that people find risk a difficult concept to understand partly because it is unfamiliar and point-probabilities are not often obvious in day-to-day life. As the structured review of risk communication formats described in section 3.2 of Chapter Three and Appendix 3.1 found, studies employing parallel statistics to assist understanding has had mixed success; generally working most effectively in highly-numerate populations (Galesic & Garcia-retamero, 2013). Providing some equivalent likelihoods as examples in the training materials might have proved useful to some respondents, however, there is also a danger of information overload leading to further confusion.

The structured review of risk communication formats also found that risk experience and perception were important factors in the communication of risk, and women drew on these aspects when trading-off the attributes ‘probability of detecting a cancer’ and ‘risk of unnecessary follow-up’. Concern about breast cancer was also a significant preference-class covariate in the quantitative results of Chapter Four. There was some evidence of an optimism bias where by women, even those with a family history, perceived themselves to be at lower than average risk. This is in line with other studies which have looked at perceptions of risk (Weinstein, 1984). Katapodi et al. (2004) reviewed 42 studies examining women’s perceived risk of breast cancer and comparing it with an objective estimate. The authors found that not only were they inaccurate; many women were conservative in their estimations. Extensive piloting to ascertain whether an optimism bias exists (or even a reverse tendency to overestimate) would be useful in order to develop a DCE tailored to the particular case study. Attempts to reduce the optimism bias have had limited success (Weinstein, 1998).

5.7.2 Strengths and limitations
The study represents an original empirical piece of work. Whilst other studies have sought to understand how people understand risk and how framing can affect its interpretation (see key examples: Henneman et al. 2013; Sprague et al. 2012; Garcia-Retamero & Cokely 2011), and qualitative research methods have been used to explored women’s perception of the risks and benefits of breast screening (Hersch et al., 2013), none have done so in a DCE setting. Similarly, a few studies have used think-aloud methods to gain a deeper-understanding of respondents’ choices in a DCE setting (Cheraghi-Sohi et al., 2007; Ryan et al., 2009), however, this is the first empirical study which has used qualitative research to understand how people balance risks and benefits in a DCE setting.
Criticisms of interviews and in particular the think-aloud method, highlight that the method relies solely on what people say, rather than what they really think (Boren & Ramey, 2000). Although an interview schedule was designed (see section 5.4.1.2) to limit sources of bias as much as possible, other limitations such as researcher-presence biases (reflexivity), nonveridicality and reactivity could have influenced the study.

The nature of qualitative analysis means there is no ‘right’ answer. Instead the approaches taken in this study have tried to be as objective as possible using a clear coding framework, consistently in NVivo® software. The processes of coding and the details of theme development are complex and cannot be articulated in their entire. The details presented in the previous sections demonstrate that the process was transparent and many steps were taken to minimise sources of bias. However, the qualitative analysis in this study was conducted solely by CV and, although extensive discussions with the supervisory team took place, this is acknowledged as a limitation. Quantitative research is not purely objective in its nature either. In choosing to quantify something, the choice of important variables or the nature of what is rational/irrational is imposed, subjectively, by the researcher. As a consequence, it is inevitable in both qualitative and quantitative research that the researcher will employ some ‘common-sense’ that will influence study progression and, as a consequence, the results.

Silverman (1994) describes the notion of ‘romanticism’ as the influence of society and culture on the subject that is almost impossible for the interviewee or respondent to remove. As many women reported seeing media articles and public health campaigns for breast cancer, it is possible that society’s view on the subject could have influenced theirs. Romanticism is distinctly different from researcher bias where the respondent is purposefully disguising their true feelings to align themselves with perceived ‘social norms’, and is instead a subconscious attention to the subject matter through indirect routes. Additionally, the inclusion of a cost attribute may have made the topic politically- and socially-sensitive. Whilst respondents were encouraged to think of costs as out-of-pocket expenses, there were inevitably some views about the preservation of existing services and affordability for the interviewee.

These sensitivities were not necessarily problems of the qualitative study; they are findings of people's thought patterns when they are completing a risky-DCE. Therefore they are acknowledged, instead, as considerations when thinking about the results. The susceptibility of this study to these sensitivities were considered before embarking on the interviews but, as discussed in Chapter Three, there are many reasons to pursue breast screening as a case study. It was felt that the familiarity of breast cancer allowed a greater
investigation of women’s thoughts of the choice task and risk attributes. Choosing a lesser known condition would likely result in other issues (such as increased cognitive burden from clinical explanations). Therefore these limitations are recognised and mentioned in both the analysis and results section of this chapter.

The recruitment of women to this study was constrained by resources and time pressures; however, the sample was varied in terms of key characteristics (age, employment). The interviewees in this study self-selected into breast screening research so for a woman to see an advertisement and respond shows some degree of inherent altruism. A selection of women who were particularly caring could have been interviewed. In addition, these women might have a pre-existing interest in breast cancer which could have, even subconsciously, influenced their choices or behaviour. Similarly, the study relied on women having free time to become actively engaged in an academic study which could have induced some social-class bias. However, the advertisements were placed in many locations and worded to attract all women from a range of backgrounds.

For these particular research questions, think-aloud provided many benefits over other qualitative research methods; most notably the verbalisation of current cognitive processes and concurrent perceptions revealed thoughts in the working memory. The method does not rely on any recall data, which is beneficial given the disadvantages associated with retrospective methods (Lundgrén-Laine & Salanterä, 2010). A serious concern is whether the act of thinking-aloud disturbs the natural decision-making behaviour and therefore generates data that is not the ‘usual’.

It is also acknowledged that the transcripts generated by the interviews contained many pauses and mumbles and some respondents required repeated prompting in the experiment. This could have been generated by two sources: 1) respondents struggling to complete the DCE task; and/or 2) respondents struggling to verbalise or communicate their thoughts. The combination of an unusual survey, about a subject they might not be familiar with, and the additional complication of verbalising their thoughts are important limitations of the think-aloud method. The warm-up exercise attempted to reduce the presence of these issues by introducing the think-aloud concept at the start of the interview and allowing the interviewee to ‘practice’ before the real task.

However, even the most articulate respondent might not be able to say their thoughts out-loud. Some processes are not easy to verbalise, and there is a certain amount of subconscious input and impulse in anyone’s behaviour (Conrad & Blair, 2009). As a result, in theory, respondents could fictionalise in response to prompts. This problem of
‘reactivity’ is not unique to the think-aloud method; any interviews encouraging someone to articulate something they find difficult are susceptible to these biases (Leow & Morgan-short, 2004; Russo et al., 1989).

The completion time for the DCE was longer than the online study indicating that verbalisation was causing some change in behaviour. Whilst the respondents might just have been ‘slowing-down’, it could also be encouraging them to consider and deliberate over items that they would otherwise have missed and might not truly reflect typical DCE behaviour. This phenomenon is referred to as ‘nonveridicality’ and is defined by Russo et al., (1989): “a protocol is considered nonveridical if it does not accurately reflect the underlying primary process” (p.760).

These limitations do not discount the findings of this qualitative study. Despite the cognitive burden of both the DCE task and the think-aloud task, there is evidence to suggest that this additional thinking actually improves the generated qualitative data. Taylor & Dionne (2000) recommend that tasks should not be too simple because anything that is familiar or involve routines will not be noticed by the short term memory and respondents will therefore not verbalise their thoughts. Instead, they actually recommend tasks which involve deliberation and a final goal (or choice, in this case) as they believe that this will generate rich and insightful data: “tasks should be novel and moderately difficult, so as to elicit conscious processing, but not so difficult as to stymie reporting” (p.415). Further research is required to investigate respondents’ choice making behaviour and reduce (if not entirely eliminate) some of the limitations mentioned above.

### 5.8 Conclusion

This study has helped to understand more about how DCE respondents make choices and trade-off risk attributes. It addressed the key research questions by revealing what people report thinking about in a DCE and how risk attributes are seen and interpreted by DCE respondents. The risk communication format did not substantially alter women’s accounts of their choice making but it did appear to relieve some of the cognitive burden in the initial choice sets by aiding the visualisation of risk.

However, qualitative research methods, in particular think-aloud, are not without their limitations. There are important considerations for the interpretation and explanation of these results. The following chapter explores an alternative method, eye-tracking, which is limited in other assumptions but helps to overcome some of the issues (namely reactivity, nonveridicality and reflexivy) identified with the qualitative research presented in this study.
Chapter Six
Benefit-risk trade-offs for breast screening: an eye-tracking study

6.1 Introduction
In the previous chapter, Chapter Five, a qualitative study using the think-aloud method was conducted, where respondents completed a DCE whilst verbalising their thoughts. This attempted to understand how the format of risk may affect respondents’ choices or decision-making strategies. However, this qualitative study raised a number of limitations with the think-aloud method. There was evidence to suggest that the act of speaking concurrently may have distracted the DCE respondents. Furthermore, the effort to fully verbalise decision strategies could have changed the way the respondents attended to the attributes. Eye-tracking is an approach which may remove some of these issues, providing a useful alternative, or supplementary, method to understand respondents’ decision-making processes through the collection of ‘objective’ (and countable) data.

Eye-tracking could offer some advantages over qualitative research methods in understanding how respondents complete a DCE. As the data provided is not textual, it requires less subjective interpretation. The qualitative study also relied on the respondent’s account of their actions. Instead, quantitative eye-tracking data can reveal information about a respondent’s overt attentional processes. Eye-tracking data has been used in many areas of research from neuroscience and psychology to computer science and marketing (Duchowski, 2002; Bialkova & van Trijp, 2011).

This chapter presents an investigation into how eye-tracking can be used to understand respondents’ attention to information presented in a DCE. The chapter first outlines the research questions to be answered in section 6.2. Section 6.3 describes eye-tracking as a method with reference to the technology’s history, its foundations in psychology, and the methods used in published studies. In section 6.4, the psychological theories described in section 6.3 are used to refine the research questions of 6.2 and develop specific hypotheses to be tested. In section 6.5, an empirical study involving an eye-tracking experiment with the DCE designed in Chapter Three is then presented, with the results and discussion described in sections 6.6 and 6.7, respectively.

6.2 Research questions
The study sought to address four key research questions which built upon the overall aim of the thesis (which was to understand if, and how, risk communication formats affect DCE valuations). Eye-tracking was used to inform understanding of the decision strategies,
rather than the choice component, of the valuation. The key research questions addressed in this study were:

1) Does DCE respondents’ attention to attributes differ with the risk communication method?
2) To what degree does subjects’ retrospective account of attribute non-attendance (ANA) match their visual attention in the DCE?
3) Does the way in which respondents seek information in a DCE differ between framing methods?
4) How does respondents’ cognitive burden differ between risk communication formats?

The research questions were answered using different data, collected using an eye-tracking study with the DCE designed in Chapter Three.

6.3 Introduction to eye-tracking as a method
In order to answer the research questions, different eye-tracking data were collected. The following sections describe how these data were collected, with section 6.3.1 explaining how the technology works based on the physiology of the eye. In section 6.3.2, the key terms and concepts related to eye-tracking are explained and defined. Section 6.3.3 presents a description of the key psychological theories which support eye-tracking as a scientific method, and explains how different measures can proxy different behaviours. In section 6.3.4, existing literature using eye-tracking methods to investigate choice is summarised.

6.3.1 Eye physiology and key concepts
The eye is pulled both horizontally and vertically by extraocular muscles, which allow fixating and tracking of stimuli even when the head remains still (Purves et al., 2001). When the muscles supporting the eye contract, the focus of gaze shifts in a movement called a ‘saccade’ (Rayner, 2009). Although the eye is never completely still, relatively short saccades or ‘jitters’ are called ‘fixations’ (Rayner, 1998). Most eye movement data come under the broad classification of either saccades or fixations (Purves et al., 2001).

Saccades are easily identifiable as the eye moves quickly in response to or in search of visual ‘stimuli’ or objects of interest. Saccadic behaviour rarely indicates information processing as the movements are so rapid that the brain is unable to consciously realise everything that is scanned, a process known as ‘saccadic suppression’ (Rayner, 1998). Instead, saccades most often represent a search for information (Kowler et al., 1995).
Saccades are distinctly different to ‘micro-saccades’, which are involuntary movements whilst an individual is attempting to fixate, and the involuntary movements which occur when an individual blinks (Otero-Millan et al., 2014). Saccades are usually described by the related variables of speed, amplitude and direction (van Beers, 2007).

What constitutes a fixation varies from study to study and is dependent on the stimulus presented (Holmqvist et al., 2011). For example, a familiar picture may be processed quicker than text, and a new diagram may be somewhere in between. Although complex algorithms exist for the identification of fixations in eye-tracking data (Salvucci & Goldberg, 2000), most studies define a threshold for a fixation as a less than one degree of movement (a measure of distance) for between 50 to 200 milliseconds (Manor & Gordon, 2003). Aggregation of the total time spent fixating, including recurrent fixations, is defined as the ‘dwell time’ to a stimulus.

Eye-tracking data provide a highly detailed record of all the locations that a user has looked at, and reducing these data to a level that can be easily analysed is challenging. One possible approach in the analysis of eye-tracking data involves segmenting coordinates to defined ‘areas of interest’ (AOI) (Vansteenkiste et al., 2014). AOI can be defined either prior to the experiment or post-experimentally once eye-movement data have been collected. Another approach to reducing the data is the generation of a ‘scan path’ describing the overall sequence of movements in terms of both saccades and fixations of a respondent, either imposed on a background image of the stimulus or as a colour-coded heat map (Holmqvist et al., 2011).

6.3.2 Technology and equipment

For many years, researchers in psychology have used eye-tracking to investigate cognitive processes (Rayner, 1998). In its most basic form, the research has involved researcher-individual observation of a participants’ eyes and manual notes on pupil dilation (Kahneman, 2012). However, more sophisticated methods have since developed in line with changes to, and availability of, technology. In the 1950s, magnetic search coils were used to track people’s eye movements which involved placing two coils on the eye, with one circling the iris on a contact lens (Duchowski, 2007). Nowadays, most eye-tracking involves less invasive equipment, commonly with a camera recording data on a computer and complex algorithms to calculate the location of the individual’s gaze (Rayner, 2009).

To track eyes, almost all modern devices record the corneal reflection on a camera positioned towards the individual’s pupil (Holmqvist et al., 2011). The corneal reflection is a glint, usually in the iris, which allows the machine to calculate the direction of the gaze.
using the distance from: 1) the camera to the eye and; 2) the eye to the screen. From the corneal reflection, the X and Y (horizontal and vertical) coordinates, which provide the location of current focus on the screen, are then recorded. The number of times this is logged a second is referred to as the speed of the tracker (Duchowski, 2007). As the eye moves from one position to another, the magnitude of the movement is measured in visual degrees (θ), rather than millimetres, as studies may involve moving stimulus and so the distance between eye and object would change. In a fixed setting, a typical computer monitor of between 17 to 20 inches has a width of 20 to 30 visual degrees (Raney et al., 2014). Figure 6.1 describes how visual degrees (θ) are measured, and how the X and Y coordinates provide a position of gaze.

**Figure 6.1: Measurement of visual degrees**

Eye-trackers are usually distinguished by their speed and, as a general rule, a good eye-tracker has a high sampling frequency and high resolution camera (Holmqvist et al., 2011). A higher sampling frequency allows a more accurate estimation of the fixation duration, as the start of the fixation is revealed earlier and the end revealed later. There is a consensus in the literature that a sampling frequency of 500 Hz is sufficiently powerful to accurately determine fixations and saccades, although anything higher is beneficial (Raney et al., 2014). An eye-tracker equipped with a high resolution camera will be able to detect the corneal reflection more easily and more accurately. Another determinant of a good eye-tracking device is its ‘latency’, which is the time taken to for the computer to make a recording. A substantial volume of processing from headset to screen to recording is required, and for some devices there is a measurable delay in this process (Holmqvist et al., 2011).
It is crucial that the eye-tracker is calibrated for each individual to ensure the eye-tracker is recording correctly (Nyström et al., 2013). The calibration procedure involves collecting fixation data from simple points on the screen in order to ascertain the true gaze position of the individual before the experiment begins (Holmqvist et al., 2011). The points are often shown as dots or crosses which move around the screen whilst fixation data are collected. A test of the calibration can be conducted by re-running the sequence and comparing the secondary fixations to the tracker’s prediction based on the first calibration data.

The calibration should involve points in all corners of the screen to ensure that the tracker is able to record in all areas (Holmqvist et al., 2011). In the corners and edges, the corneal reflection can disappear, which therefore invalidates the computer's calculations as well as result in missing data. Similarly, for individuals with visual aids (glasses, contact lenses) or heavy eye-make-up, the far corners can often induce another reflection which may confuse the recording and create anomalous data.

Tracking devices can either record both eyes (binocular) or a single eye (monocular) (Holmqvist et al., 2011). When both eyes are recorded, an average of the horizontal and vertical coordinates from each eye are taken. However, most people generally have an ‘active’ and ‘lazy’ eye and literature suggests that the active, dominant eye should only be tracked (Nyström et al., 2013). If a participant performs poorly in the calibration, then an alternative eye should be tried.

There are three broad categories of modern eye-tracking devices: 1) head mounted; 2) remote; and 3) head-supported towers. Head mounted eye-trackers, such as smart glasses or helmet cameras, offer participants some freedom, but these are harder to calibrate and can be cumbersome to wear. These eye-trackers are often used to understand how objects are attended to in a dynamic situation, for example, whilst the participant is engaged with a shopping activity (Holmqvist et al., 2011). Remote eye-trackers involve no head restraint and instead let the participant move freely, with algorithms used to detect non-eye movements (Bohme et al., 2006). However, the additional calculations to distinguish head and eye movements are a burden to the processing capacity of the computer, and generally result in a lower frequency and, as a consequence, have decreased precision (Holmqvist et al., 2011).

Head-supported towers involve the use of a forehead and chin-rest. Whilst being contactless, these can be uncomfortable and unnatural for some participants. These devices are also often immobile, due to their heavy processing power, and require stability of the head because of their high frequency. However, head-supported towers are the most accurate
and precise equipment available for researchers. For studies where the individual is not required to move and the stimuli are stationary (such as a survey), a head-supported tower eye-tracker provides the best quality data (Holmqvist et al., 2011; Raney et al., 2014).

6.3.3 Psychological foundations
Cognitive processes are incredibly complex and it may be impossible to measure what any individual is actually thinking at any one time. However, psychologists have studied eye-movements in order to understand more about people’s brain functioning and how information is processed (Rayner, 1998). The ‘eye-mind hypothesis’, suggested by Just & Carpenter (1980), provided the underpinning to most psychological analyses of eye-tracking data. The following sections describe the three key behaviours expanded to answer the research questions stated in section 6.2. The theoretical foundations supporting visual attention, as acquired through fixation data, are described in section 6.3.3.1. In section 6.3.3.2, information searching behaviour as exhibited through fast, saccadic movement, are described. In section 6.3.3.3, ‘cognitive pupillometry’, a measure of task difficulty through pupil dilation, is described.

6.3.3.1 Visual attention
The logic of using eye-tracking data as criterion data is based on a generally accepted assumption that where people look is indicative of their thoughts. Just & Carpenter (1980) assertively stated that: “there is no appreciable lag between what is fixated and what is processed” (p.331). Therefore attention, as measured in fixations, can be thought of as a quantification of an individual’s information processing. Although it could be argued that individuals who are day-dreaming may inadvertently fixate on something whilst their thoughts are somewhere else, this behaviour is often shown through attention to areas of white-space with no information, or off the screen entirely (Rayner, 1998). This dreaming behaviour is also known as ‘covert attention’ and is distinctly different to information-acquiring fixations to stimuli (Henderson et al., 1989).

6.3.3.2 Information searching
Data generated between fixations are generally classified as saccades. These rapid movements occur when individuals are searching for information but are not taking account of the stimulus they are scanning (Kowler et al., 1995). Saccadic patterns in eye-tracking data have been used to explain visual responsiveness, such as how quickly an individual can find an answer (Findlay, 2009). In the context of choices, saccades have been used to understand how individuals seek information to make a decision, with research suggesting vertical movements in line with EUT (Arieli et al., 2011, 2009).
6.3.3.3 Cognitive pupillometry

In addition to eye movements, some eye-trackers have the capacity to collect other data on the participant’s eye. Pupillometry is the measurement of the pupil, and cognitive pupillometry refers to the change in pupil size due to task burden (Hartmann & Fischer, 2014). When individuals are thinking hard about something which is difficult or requires significant memory load, the pupils may dilate (Laeng et al., 2012). In early studies, this was an incidental finding discovered when observing participants in laboratory tasks (Kahneman, 2012). However, there are now many data-driven studies which have empirically demonstrated a relationship between task complexity and pupil dilation (Binda et al., 2014; Beatty & Wagoner, 1978).

6.3.4 Research to date

Whilst eye-tracking has grown as a method, and is frequently used by psychology departments in academic institutions (Rayner, 1998), there are very few examples in either health economics or economics. The results of the systematic review presented in Appendix 5.1 did not identify any published healthcare DCEs which had used eye-tracking methods.

A DCE published in 2015, that elicited preferences for food consumption, used eye-tracking data to investigate visual ANA as measured by the number of times respondents re-attended attribute levels in a choice set (see Balcombe et al. (2015)). A study by Bialkova & van Trijp (2011) looked at respondents’ attention to food labelling in a choice survey set up in a similar way to a DCE. In this study, the analysis lacked detail but the authors used fixation and response-time data to analyse the effect of different labelling of yoghurt packages on consumer choice. Despite parallels in these choice-based eye-tracking studies, neither looked at a risk attribute in a DCE, and neither involved a healthcare good or service.

The next section of this chapter, section 6.4, presents an empirical study with an innovative application of eye-tracking to the DCE designed in Chapter Three.

6.4 Hypotheses

Section 6.3 identified outcomes of eye-tracking data that potentially have value in this PhD. Linking the psychological theories and research questions of section 6.2, four hypotheses were generated, each matching the respective research question.

Hypothesis one: The ‘eye-mind hypothesis’ introduced in section 6.3 suggested attention to stimulus indicates information is processed by the participant. Therefore, an individual
completing a DCE and attending an attribute more (longer) would be conducting more information processing resulting in a better, or more informed, choice. Therefore it was hypothesised that participants presented with the IAP version of the DCE would pay more attention (and make more informed choices) to the risk attributes in comparison to those who received the percentages only. Visual attention was measured by: 1) number of fixations; 2) longer fixations; and 3) in total, a longer dwell time.

*Hypothesis two:* Individuals completing a DCE who reported not attending an attribute would pay less visual attention to these attributes. Attention to an attribute was measured as total fixation duration (dwell time) to the attribute.

*Hypothesis three:* The communication of risk could affect the way respondents’ searched for information in order to make their choice in the DCE.

*Hypothesis four:* It was hypothesised that women who received this risk communication format would have reduced pupil dilation in the task as these would aid choice-making and make the task less cognitively burdensome.

### 6.5 Methods

The following sections describe the methods used in an eye-tracking experiment, where respondents completed the DCE survey designed in Chapter Three whilst their eyes were monitored.

#### 6.5.1 Recruitment

The sample selection criteria were limited to females, fluent in English and between 18 and 70 years of age (for reasons discussed in Chapter Three). As stated previously in section 3.4.1 of Chapter Three, 70 years was chosen as a maximum age as this is the current cut-off for routine screening in England (Independent UK Panel on Breast Cancer Screening, 2012). No other sample restrictions were included as initial discussions with breast cancer experts in the design of the DCE revealed that they were interested in the views of all women (see Chapter Three). In light of the expert advice, both the internet panel (see Chapter Four) and the qualitative study (see Chapter Five) sampled women in the 18 to 70 age categories and therefore, for consistency, the sampling frame for this eye-tracking study continues with the same inclusion criteria.

The advertising strategy replicated that of the qualitative study described in section 5.4.1.1 of Chapter Five. Examples of the advertisements can be found in Appendices 6.1 and 6.2. Following the recommendations of the University’s ethics committee, responders to the advertisements were only incentivised at first contact with the researcher, when they were
informed about the £10 Amazon voucher, as per the recommendation of the ethics committee. At this first point of researcher contact, the potential participants were also provided with an information sheet containing frequently asked questions (Appendix 6.3). The advertisement responders were then asked to read the sheet, take time to think about the study and then, if they were interested, to arrange a visit to the eye-tracking laboratory located in the Zochonis Building at The University of Manchester.

The study aimed for a sample size of 40 women, on the basis of a similar choice study by Bialkova & van Trijp (2011), who also used a sample of 40 and the results of a pilot study conducted in May 2014 (see Appendix 6.4 for a description). The number of subjects was also restricted by time constraints (as the eye-tracking laboratory was shared, time using the equipment was limited) and resource constraints (if the £10 Amazon voucher thank-you was maintained). Therefore recruitment continued until 40 successful experiments had been completed.

6.5.2 Apparatus
The head-mounted EyeLink® 1000 was used in this experiment (SR Research, 2012). The eye-tracking device calculates the participant’s gaze position using a camera to detect the corneal reflection due to an infrared illuminator. The device recorded the eye position a thousand times a second, every millisecond (ms). The head rest was positioned 43 centimetres from the screen, as per the manufacturer’s recommendations, and this distance was re-measured for every participant. Whilst the machine had a capacity for binocular recording, monocular recording of the dominant eye was conducted for the reasons discussed in section 6.3.2. The experiment took place in a dark, window-less room with minimal luminosity. Choices for the DCE were made via a handheld games controller.

The training materials and choice sets of the DCE survey designed in Chapter Three was programmed for the eye-tracker with assistance from EyeLink® Experiment Builder software (SR Research, 2012). The DCE was identical to that used in Chapters Four and Five except that no video was included in the training material, as the video file was too large to be loaded into the experiment. Eye-tracking data relating to the video would have been disregarded in the analysis as data relating to a dynamic stimulus would have required completely different analytical methods. The eye-tracking DCE also included three additional warm-up questions, which allowed participants to acquire a familiarity with the handheld controller buttons. All images in both risk formats were exactly the same size to the nearest pixel. The set-up of the eye-tracking apparatus is shown in Figure 6.2.
A shows the screen which displayed the training information and choice sets. B represents the camera recording the participant’s eye. C is the infrared illuminator which allowed recording in the dark room. D points to the handheld controller that participants used to make their choice.

6.5.3 Data collection

Once a woman had arranged to visit the laboratory, she was asked again to read through the information sheet and, if she felt happy, consent to the study by signing the form shown in Appendix 6.5. The women who consented were randomly assigned (using a random number table) to receive either the PO version from design block one or two, or the IAP version from design block one or two. The participants were then asked to place their chin on the head rest, make themselves comfortable, and refrain from speaking. The calibration process then began. If the calibration was ‘good’ (the corneal reflection was consistently recorded), then it was validated through an additional re-calibration which ensured all corners of the screen were recordable before the survey began. If the calibration or validation failed, the recording eye was changed and the procedure restarted. In the event that neither eye could be calibrated, the participant was thanked for their time and informed that the experiment could not be completed.
Whilst answering the DCE, a between choice set calibration occurred called ‘drift correction’. This was to correct for any ‘drift’ which had occurred and to improve the accuracy of the collected data. Drift occurs when the position of the eye moves slightly and the calibration is no longer perfect. In the event that the drift correction indicated the participant had moved their head or the tracker had stopped recording, a complete re-calibration and validation procedure was conducted. The eye-movements were recorded for each of the eleven choice sets as separate independent ‘trials’.

Background questions were completed by the participant on an iPad after they had finished the DCE with the eye-tracker. Questions included socio-demographic details and self-reported ANA. The iPad-based survey used in this experiment can be found in Appendix 6.6.

6.5.4 Analysis
The effect of risk attribute format on a range of eye-tracking-based outcomes was investigated to answer the research questions stated in section 6.2, and test the four hypotheses presented in section 6.4. Section 6.6.4.1 describes the first stage of preparing the collected data for analysis. Section 6.6.4.2 explains how outcomes of interest were collapsed to an individual-trial level, and in section 6.6.4.3, the regressions conducted to analyse the eye-tracking data are specified.

6.5.4.1 Data preparation
The following section describes how the eye-tracking data was prepared for the analysis, with generated outcomes to answer the research questions stated in section 6.2. The first stage of data reduction, defining AOI, is explained in section 6.5.4.1.1. The key outcomes of interest: fixations, saccade and pupil size, are described in sections 6.5.4.1.2, 6.5.4.1.3 and 6.5.4.1.4, respectively.

6.5.4.1.1 Areas of interest
The first stage of analysis involved reducing the data to pre-defined AOI. This process reduced the X and Y coordinates and pupil size data into a series of variables of interest and reduced the file to a workable size. AOI reflect the participant’s attention to a particular segment of a task. These AOI quadrants were defined for each choice set on the screen, based on the sections of the task that might be stimulating for the participants. Figure 6.3 shows the AOI defined in each choice set, for example, the attribute titles, the levels presented, and the response options. Of particular interest in this study were the outlined segments; the AOI for the attribute levels. The areas outside the AOI were used to
measure the amount of gaze in the ‘white space’, and attention to this white space provided an approximate measure of validity.

**Figure 6.3: Pre-defined AOI for each choice set**

6.5.4.1.2 Fixations

Fixations were measured in milliseconds and were conservatively defined as a less than 1° movement for 75 milliseconds. This fixation duration threshold was defined based on previous studies, such as Bialkova & van Trijp (2011) who assigned a threshold of 80 milliseconds to respondents completing a choice set of images and numbers. If a fixation was under this threshold, and another fixation occurred within 1° of the original fixation, then the fixations were merged together. Merging adjacent fixations allowed identification of fixations which may have been missed due to measurement errors.

Collecting fixation data allowed analysis of information processing via the respondents’ attention to the choice task. Fixation data can be analysed in terms of: 1) the number of fixations; 2) the average length of a fixation; and 3) the total dwell time (the sum of all fixation durations) to an AOI. These three outcomes were of interest in the analysis of the eye-tracking data and can be useful in testing hypothesis one, as they each provide different measures of visual attention.

Self-reported ANA was captured as a dummy variable depending on the respondents’ answer to the iPad-based survey question (see Appendix 6.6, p.495). These data were used to compare mean dwell time to each attribute which can be thought of as measuring the
latent scale from which individuals assign ANA in the self-reported question. These data were used to test hypothesis two, which theorised respondents who report ANA would have a shorter mean dwell time to those attribute AOI.

6.5.4.1.3 Saccades

No amplitude threshold was defined for saccades and any observations that were not fixations constituted movement. Saccades were measured by their direction (angle of movement). Amplitude and velocity were not of interest in this study because they would yield little information about respondents’ decision-making strategies.

The EyeLink® 1000 tracker records degrees of direction between -180° and 180°, with 0° being a perfectly horizontal movement to the right. A rightward saccade was defined as a movement between -45° and 45°; a downward saccade was defined as a movement between -135° and -45°; a leftward saccade was defined as a movement between -135° and 135°; and a upward saccade was defined as a movement between 45° and 135°. Figure 6.4 describes each directional movement and its definition in degrees.

![Figure 6.4: Definition of direction from saccade angles](image)

Blinks were identified by the EyeLink®’s in-built software (SR Research, 2013). When a participant blinks whilst their eyes are being tracked, the gaze shifts down in what could be interpreted as a saccade. However, a blink was identifiable as it was immediately followed by a missing pupil image on the camera as the eyelid closes. These data were acknowledged but disregarded from the analysis.

The saccade data of interest was the number of saccades made, and the direction of these saccades. These data tested hypothesis three which sought to understand how respondents seek information in a choice set before selecting their preferred alternative. No specific hypothesis existed about the number of saccades a respondent may make, but more up-
down movements could suggest the choices were made in line with EUT and Lancaster’s Theory as the respondent was taking into account the alternative as a whole.

6.5.4.1.4 Pupillometry

Pupil size was calculated by the eye-tracker which counts the number of black pixels on the camera image of the eye to identify a measurement of the pupil diameter. The EyeLink® 1000 is generally regarded as a good measurer of pupil size, as alternative trackers use an ellipse fitting highly affected by noise. Pupil dilation was calculated as the difference between the minimum pupil size and maximum pupil size. This was kept as an absolute measure rather than as a percentage which can be inflated when baseline pupil size is small (Wang, 2010).

With the EyeLink® 1000, pupil size data were not calibrated, and units of pupil measurement typically vary between studies. Pupil size was recorded as an integer number, based on the number of pixels but measured in arbitrary units meaning that results cannot be compared across studies, or even within studies if there was inconsistent luminosity or the stimuli appeared at different locations on the screen, as this can affect the measure of pupil size. However, in this experiment the choice set stimuli, either percentages or icon arrays, occurred in precisely the same location on the screen and the head-mount, screen and camera were identically located for each subject. In addition, all experiments took place in the same laboratory with identical equipment set-up and light sources, maintaining the same luminosity.

Pupil size data was analysed as: 1) the average pupil size per individual fixating to an attribute in a choice-set (trial); and 2) the average change in pupil size per individual fixating to an attribute in a trial. The pupil size data were used as a measure of cognitive burden and was used to test hypothesis four. Hypothesis four suggested that larger pupils indicate a more demanding task, and a large change in pupil size could be the result of a question that is more cognitively challenging.

6.5.4.1.5 Response time

Response time was measured from the end of the drift correction to the participant pressing a controller button to make their choice, and was recorded to the nearest millisecond. A shorter response time could indicate that respondents were able to make their choices more quickly because the information was easier to acquire.

Variables derived from the background data collected from the iPad-based survey at the end of the experiment were not used as covariates in the analysis of the eye data. As
respondents were randomised to receive either risk format of the DCE, any observable differences in characteristics would have occurred by chance and there was no hypothesised-relationship between socio-demographic characteristics and visual attention.

6.5.4.2 Collapsing data to derive outcomes of interest

For most outcomes, data were available at a saccade or fixation level as made by each participant, in each choice set. To analyse the effects of risk format, the eye-tracking data of all participants completing all choice sets were collapsed to an individual-trial level. This allowed the results to be interpreted as a statistic per choice set for a respondent receiving the icon arrays or percentages only. For mean outcomes, this was achieved by taking an average of outcomes at the fixation/saccade/pupil size/response duration level for each individual in each choice-set (trial). For aggregate outcomes (dwell time) a sum of outcomes was taken. Number outcomes (number of saccades, number of fixations) were calculated through counts. For binary outcomes (direction of saccade), this was calculated as the proportion of times at the saccade level (in a trial) where the binary variable equalled unity (1).

To illustrate the collapsing of data, a highly simplified example of fixation-level data can be found in Figure 6.5. Figure 6.5 shows hypothetical-participant 1 making 9 fixations in choice set 1 and attending all attributes except the cost of alternative B. The participant attended the detect attributes 4 times, the risk attributes twice, and the cost attribute once. The participant also looked at the white space once for 81 milliseconds and the text about detection in the ‘No Screening’ option for 150 milliseconds.

Figure 6.5: Example of eye-tracking fixation data

<table>
<thead>
<tr>
<th>ID</th>
<th>Fixation duration</th>
<th>Pupil size</th>
<th>AOI</th>
<th>Choice set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78</td>
<td>1000</td>
<td>RiskA</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>123</td>
<td>1100</td>
<td>DetectA</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>85</td>
<td>1220</td>
<td>DetectB</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>196</td>
<td>1190</td>
<td>CostA</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>178</td>
<td>1400</td>
<td>DetectA</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>93</td>
<td>890</td>
<td>RiskB</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>182</td>
<td>850</td>
<td>DetectA</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>81</td>
<td>1000</td>
<td>None_detected</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>150</td>
<td>1100</td>
<td>None_detected</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>113</td>
<td>1340</td>
<td>RiskA</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>76</td>
<td>950</td>
<td>RiskB</td>
<td>2</td>
</tr>
</tbody>
</table>

The collapsed fixation data analysed would then take the form of Figure 6.6, with the key outcomes of interest presented per participant, per trial. With key variables including
number of fixations (NOF), mean fixation duration (MF), dwell time (DT), mean pupil size, change in pupil size and self-reported ANA at an individual-trial level. The average fixation duration to the detection attributes was 142 milliseconds, to the risk attributes 85.5 milliseconds and to the cost attribute was 196 milliseconds. The total dwell time was 568 milliseconds to the detect attributes; 171 milliseconds to the risk attributes; and 196 milliseconds to cost. The mean pupil size for the whole trial was 1092 and the change in size was 550.

Figure 6.6: Example of collapsed fixation data to key AOI

<table>
<thead>
<tr>
<th>ID</th>
<th>AOI</th>
<th>Choice set</th>
<th>NOF</th>
<th>MF</th>
<th>DT</th>
<th>Mean pupil size</th>
<th>Change in pupil size</th>
<th>ANA detect</th>
<th>ANA risk</th>
<th>ANA cost</th>
<th>Risk format</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Risk</td>
<td>1</td>
<td>2</td>
<td>85.5</td>
<td>171</td>
<td>945</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Detect</td>
<td>1</td>
<td>4</td>
<td>142</td>
<td>568</td>
<td>1143</td>
<td>550</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Cost</td>
<td>1</td>
<td>1</td>
<td>196</td>
<td>196</td>
<td>1190</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Risk</td>
<td>2</td>
<td>2</td>
<td>126</td>
<td>420</td>
<td>945</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Detect</td>
<td>2</td>
<td>6</td>
<td>156</td>
<td>672</td>
<td>1126</td>
<td>360</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Cost</td>
<td>2</td>
<td>2</td>
<td>78</td>
<td>158</td>
<td>1191</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Risk</td>
<td>3</td>
<td>3</td>
<td>90</td>
<td>246</td>
<td>949</td>
<td>110</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Detect</td>
<td>3</td>
<td>5</td>
<td>156</td>
<td>736</td>
<td>1104</td>
<td>442</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>Cost</td>
<td>3</td>
<td>3</td>
<td>98</td>
<td>282</td>
<td>1193</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>%</td>
</tr>
</tbody>
</table>

6.5.4.3 Regression-based analysis

The collapsed data were analysed using ordinary least squares (OLS) in a linear regression model attempting to fit the observed data by minimising the sum of the squared residuals (Wooldridge, 2008). This method was chosen to investigate the effect of the risk communication method on the outcomes of interest. For all outcomes, the regression specification in equation 6.1 was estimated separately for the detection, risk and cost attributes:

$$y_{i,c} = \alpha + \beta IAP_t + \epsilon_{i,c} \quad [6.1]$$

$y_{i,c}$ denotes the outcome for individual $i$ in choice set $c$. $IAP_t$ is a binary variable which equals zero if individual $i$ completes the DCE where risk was framed as a percentage only, and equal to unity if individual $i$ completes the DCE where risk was formatted with an additional icon array. $\epsilon_{i,c}$ is a zero mean error term, assumed uncorrelated with the method of risk formatting. Clustering of standard errors at an individual level was included to allow for observations by the same individual to be correlated. The regression of equation 6.1 produced the estimated coefficients shown in equations 6.2 and 6.3:

$$\hat{\alpha} = \bar{y}^{PO} \quad [6.2]$$
\[ \hat{\beta} = \bar{y}^{IAP} - \bar{y}^{PO} \]  

[6.3]

The estimated constant, \( \hat{\alpha} \), is the average of individual-trial level outcomes for individuals who completed the DCE with risk framed as a percentage only, \( \bar{y}^{PO} \) and \( \hat{\beta} \) is the average difference in average outcomes between individuals who completed the DCE with risk attributes formatted as both a percentage and icon array (\( \bar{y}^{IAP} \)) and individuals who completed the DCE with risk communicated as a percentage only (\( \bar{y}^{PO} \)).

To study the effects of self-reported ANA, the outcome was measured as the mean dwell time to an attribute. To estimate this effect, the regression specification in equation 6.4 was estimated separately for the detection, risk and cost attributes:

\[ y_{i,c} = \alpha + \beta \text{ANA}_i + \varepsilon_{i,c} \]  

[6.4]

\( y_{i,c} \) denotes the total fixation duration for individual \( i \) in trial \( c \). \( \text{ANA}_i \) is a binary variable which equals zero if individual \( i \) reported that they attended to an attribute, and equal to unity if individual \( i \) reported not attending to an attribute. \( \varepsilon_{i,c} \) is a zero mean error term, assumed uncorrelated with self-reported ANA. The regression of equation 6.4 produced the estimated coefficients of 6.5 and 6.6:

\[ \hat{\alpha} = \bar{y}^{A} \]  

[6.5]

\[ \hat{\beta} = \bar{y}^{ANA} - \bar{y}^{A} \]  

[6.6]

The estimated constant, \( \hat{\alpha} \), is the mean dwell time per attribute per trail for individuals who reported attending an attribute (\( \bar{y}^{A} \)), and \( \hat{\beta} \) is the difference in average (mean) dwell time between individuals who reported not attending to an attribute (\( \bar{y}^{ANA} \)) and individuals who reported attending to an attribute (\( \bar{y}^{A} \)).

Table 6.1 summarises the different types of eye-tracking data, how the outcomes are defined and their units of measurement, the meaning of \( \bar{y} \) and \( \hat{\beta} \) for each OLS regression, and the interpretation of the regression results.
Table 6.1: Outcomes of interest in the eye-tracking data

<table>
<thead>
<tr>
<th>Label</th>
<th>Outcome</th>
<th>Definition</th>
<th>Type</th>
<th>Unit</th>
<th>(\hat{y}) in OLS regression</th>
<th>(\hat{\beta}) in OLS regression</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment</td>
<td>Response time (RT)</td>
<td>Time between choice set appearing and choice being made</td>
<td>Mean</td>
<td>Milli-seconds</td>
<td>Mean RT from starting a trial to pressing a button</td>
<td>Difference in mean RT when risk is presented with an icon array</td>
<td>Hard experiments take longer to complete</td>
</tr>
<tr>
<td>Fixation</td>
<td>Number of fixations (NOF)</td>
<td>Number of separate occasions an individual fixates on information</td>
<td>Count</td>
<td>NOF</td>
<td>Mean NOF by an individual to an attribute in a trial</td>
<td>Difference in mean NOF when risk is presented with an icon array</td>
<td>More fixations mean more information processing</td>
</tr>
<tr>
<td>Fixation</td>
<td>White space fixations</td>
<td>Number of separate occasions an individual fixates on areas containing no information</td>
<td>Count</td>
<td>NOF</td>
<td>Mean NOF by an individual to a white space in a trial</td>
<td>Difference in mean NOF when risk is presented with an icon array</td>
<td>More fixations to white space means confusion/ no interest</td>
</tr>
<tr>
<td>Fixation</td>
<td>Average fixation duration</td>
<td>Mean time of a fixation</td>
<td>Mean</td>
<td>Milli-seconds</td>
<td>Mean duration of a fixation to an attribute by an individual in a trial</td>
<td>Difference in mean duration when risk is presented with an icon array</td>
<td>Longer fixations mean more information processing</td>
</tr>
<tr>
<td>Fixation</td>
<td>Total dwell time (DT)</td>
<td>Total time of all fixations</td>
<td>Sum</td>
<td>Milli-seconds</td>
<td>Mean DT to an attribute by an individual in a trial</td>
<td>Difference in mean DT when risk is presented with an icon array</td>
<td>Longer DT means more information processing</td>
</tr>
<tr>
<td>Fixation and ANA</td>
<td>Total dwell time (DT)</td>
<td>Number of separate occasions an individual searches for information information seeking</td>
<td>Sum</td>
<td>Milli-seconds</td>
<td>Mean DT to an attribute by an individual in a trial</td>
<td>Difference in mean DT when risk is presented with an icon array</td>
<td>Self-reported ANA should mean shorter DT as attributes ignored</td>
</tr>
<tr>
<td>Saccade</td>
<td>Number of saccades (NOS)</td>
<td>Number of separate occasions an individual searches for information</td>
<td>Count</td>
<td>NOS</td>
<td>Mean NOS by an individual in a trial</td>
<td>Difference in mean NOS when risk is presented with an icon array</td>
<td>No hypothesis</td>
</tr>
<tr>
<td>Saccade</td>
<td>Saccade direction</td>
<td>Direction of eye-movements in information seeking</td>
<td>Count</td>
<td>Binary (0=up-down; 1=left-right)</td>
<td>Proportion of saccades to an attribute in a trial, which are left-right</td>
<td>Difference in mean proportions when risk is presented with an icon array</td>
<td>More up-down movements mean more whole-good calculations</td>
</tr>
<tr>
<td>Pupil</td>
<td>Average pupil size</td>
<td>Mean pupil size in a fixation</td>
<td>Mean</td>
<td>Pixel-units</td>
<td>Mean pupil size of an individual when fixating to an attribute in a trial</td>
<td>Difference in mean pupil size when risk is presented with an icon array</td>
<td>Larger pupil size indicates more cognitive burden</td>
</tr>
<tr>
<td>Pupil</td>
<td>Pupil dilation</td>
<td>Change in pupil size whilst fixating (largest-smallest)</td>
<td>Mean</td>
<td>Pixel-units</td>
<td>Mean difference in pupil size of an individual when fixating to an attribute in a trial</td>
<td>Difference in mean pupil size change when risk is presented with an icon array</td>
<td>Larger changes in pupil size indicates more cognitive burden</td>
</tr>
</tbody>
</table>
6.6 Results
The results section starts with a description of the participants’ responses and the validity of the experiment in sections 6.6.1 and 6.6.2, respectively. The results of the regressions relating to the visual attention and ANA are presented in sections 6.6.3 and 6.6.4, respectively. The results relating to the saccades made by participants are described in section 6.6.5 and the pupil size regression results are presented in section 6.6.6.

6.6.1 Participants
In total, 42 women were recruited, although two women were excluded because their eye movements could not be accurately recorded due to calibration failure. In total, forty women aged between 18 and 63 years old were included. Twenty women were randomised to each risk version of the DCE. Within each version, ten women were randomised again to each of the two design blocks.

The complex recruitment strategy confounded with women from the qualitative interviews presented in Chapter Five makes it difficult to attribute a precise response rate for this particular study. For more information on the recruitment results, see Appendix 5.4.

The whole experiment, including the calibration, choice questions, and the completion of background questions on the iPad, took approximately 20 minutes. The average response time for a single choice was 13388 milliseconds (13.4 seconds), as shown in Table 6.2. This was slightly shorter (half a second) for the IAP format; however, the difference was not statistically significant.

Table 6.2: Mean response time in milliseconds from starting the choice set to pressing a button, and differences between risk formats

<table>
<thead>
<tr>
<th>Risk format</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO % [%α]</td>
<td>13387.67 ms (685.21)</td>
</tr>
<tr>
<td>IAP ↑ icon array difference [%β]</td>
<td>-444.82 ms (968.54)</td>
</tr>
</tbody>
</table>

Observations 440

*Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001*

6.6.2 Experimental validity
Figures 6.7 and 6.8 show examples of the eye-movements of a respondent from the IAP and PO DCEs, respectively. A visual examination of the scan paths of participants indicated that the eye-tracking experiment had face validity, with almost all fixations to the
pre-defined AOI. Table 6.3 shows that, on average, each participant made only two
fixations to areas of white space (with no information) per choice-set question. There was
no significant difference between the two risk formats.

Table 6.3: Mean number of fixations to white space per choice set, and differences between risk
formats

<table>
<thead>
<tr>
<th>Risk format</th>
<th>Mean number of white space fixations</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO % [α]</td>
<td>2.379 (0.68)</td>
</tr>
<tr>
<td>IAP †</td>
<td>2.403</td>
</tr>
<tr>
<td>Icon array difference [β]</td>
<td>0.024 (0.02)</td>
</tr>
<tr>
<td>Observations</td>
<td>440</td>
</tr>
</tbody>
</table>

Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001

6.6.2.1 DCE validity

Results of heteroscedastic conditional logit models using choice data from the eye-tracking
participants can be found in Appendix 6.7. All coefficients possessed the expected sign
(negative coefficients on the attributes ‘risk of unnecessary follow-up’ and ‘cost of
screening programme’; and a positive coefficient on the attribute ‘probability of detecting a
cancer’). However, the coefficient on the cost attribute was statistically insignificant. Two
participants failed the internal validity check (a failure rate of 5%).
Figure 6.7: Example scan path from a respondent completing the DCE with icon arrays and percentages

Figure 6.8: Example scan path from a respondent completing the DCE with percentages only
6.6.3 Visual attention: fixation and dwell times

Hypothesis and research question one sought to understand if, and how, visual attention differed by risk framing communication format through an examination of the effects on fixations. As shown in Table 6.4, the risk and detection attributes had the highest number of fixations, attracting almost twice as many fixations as the cost attribute. The average number of fixations made when looking at a choice-set was not significantly different between the two risk versions of the DCE.

Table 6.4 also shows that not only did DCE respondents fixate more to the risk and detection attributes, the fixations to these attributes were also longer in duration. Although respondents made more fixations to the risk attribute, the mean fixation time was shorter. There was no significant difference in fixation duration between the risk communication formats.

When the number of fixations and the fixation duration were aggregated to measure the complete dwell time to an attribute, results were maintained. The eye-tracking participants spent a total of 1.8 seconds, 1.9 seconds and 0.8 seconds on the attributes detect, risk and cost, respectively. Eye-tracking participants who received the IAP spent longer processing all attributes, although the difference was not significant.
Table 6.4: Visual attention to each attribute in a trial (hypothesis one)

<table>
<thead>
<tr>
<th>Risk format</th>
<th>Attribute</th>
<th>Probability of detecting a cancer</th>
<th>Risk of unnecessary follow-up</th>
<th>Lifetime cost of screening programme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean number of fixations to each attribute in a trial, and differences between risk formats</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO% [α]</td>
<td>7.605</td>
<td>8.500</td>
<td>4.045</td>
<td></td>
</tr>
<tr>
<td>IAP †</td>
<td>8.137</td>
<td>9.832</td>
<td>4.400</td>
<td></td>
</tr>
<tr>
<td>Icon array difference [β]</td>
<td>0.532</td>
<td>1.332</td>
<td>0.355</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk format</th>
<th>Mean duration of a fixation to each attribute in a trial in milliseconds, and differences between risk formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO% [α]</td>
<td>237.267 ms (14.95)</td>
</tr>
<tr>
<td>IAP †</td>
<td>261.727 ms</td>
</tr>
<tr>
<td>Icon array difference [β]</td>
<td>24.460 ms (20.61)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Risk format</th>
<th>Mean dwell time (total duration of fixations) to each attribute in a trial in milliseconds, and differences between risk formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO% [α]</td>
<td>1789.427 ms (219.59)</td>
</tr>
<tr>
<td>IAP †</td>
<td>1982.659 ms</td>
</tr>
<tr>
<td>Icon array difference [β]</td>
<td>193.232 ms (275.11)</td>
</tr>
</tbody>
</table>

Observations 440

*Standard errors in parentheses, *p<0.05, **p<0.01, ***p<0.001

### 6.6.4 Attribute non-attendance

Hypothesis and research question two sought to understand if, and how, self-reported ANA matched ANA in the eye-tracking data through a comparison of fixations. When the eye-tracking dwell-time data were compared with self-reported ANA, it was found that the mean dwell time was significantly lower for people who reported non-attendance to these attributes (for’ probability of detecting a cancer’ and ‘risk of unnecessary follow-up’). The results in Table 6.5 shows the difference was considerable, with non-attendance to risk resulting in a 25% lower dwell time to the risk attributes, and non-attendance to detect resulting in over 40% shorter dwell times. However, there was no statistically significant difference between dwell-times to the cost attribute between those who reported attending it and those who stated ANA.
Table 6.5: The effect of self-reported ANA on total dwell-time per trial in milliseconds (hypothesis two)

<table>
<thead>
<tr>
<th>ANA</th>
<th>Probability of detecting a cancer</th>
<th>Risk of unnecessary follow-up</th>
<th>Lifetime cost of screening programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>No self-reported ANA [α]</td>
<td>1970.293 ms (145.70)</td>
<td>2010.305 ms (203.04)</td>
<td>881.676 ms (113.22)</td>
</tr>
<tr>
<td>Self-reported ANA</td>
<td>1127.796 ms (231.37)</td>
<td>1443.909 ms (203.04)</td>
<td>808.900 ms (344.48)</td>
</tr>
<tr>
<td>Difference [β]</td>
<td>-842.497 ms*** (-231.37)</td>
<td>-566.396 ms** (-203.04)</td>
<td>-72.776 ms (-344.48)</td>
</tr>
</tbody>
</table>

Observations 440

Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001

In this study, there was complete visual attention to the attribute ‘probability of detecting a cancer’ by participants who received the PO version. However, there were some participants who did not attend to an attribute at all in some choice sets, exhibiting complete visual ANA. As Table 6.6 shows, when risk was framed as a percentage only, complete non-attendance to the risk attribute occurred in seven choice-sets, compared to only once when risk was framed with an icon array. Table 6.6 also shows that in terms of complete visual ANA, the cost attribute was most neglected by the participants with 24 choice sets completed with no visual attention to the attribute, at all.

Table 6.6: Number of choice-sets with no attention to an attribute for each risk format

<table>
<thead>
<tr>
<th>Risk format</th>
<th>Detect</th>
<th>Risk</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAP</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>(0.23%)</td>
<td>(0.23%)</td>
<td>(3.41%)</td>
</tr>
<tr>
<td>PO %</td>
<td>0</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>(0.00%)</td>
<td>(1.59%)</td>
<td>(2.05%)</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>8</td>
<td>24</td>
</tr>
</tbody>
</table>

As a percentage of all choice-sets in parentheses

6.6.5 Saccades

Hypothesis and research question three sought to understand if, and how, respondents searched for information in a choice-set differed by risk communication formats.

Information searching was evaluated through analysis of the saccade data, with more upward-downwards eye-movements in line with EUT and Lancaster’s Theory. There was no significant difference between the two risk communication formats in terms of the average number of saccades a participant made in each trial (around 48 movements). When it came to the direction of the saccade, participants completing the DCE in both risk versions made more horizontal (left-right) saccades than vertical (up-down). As table 6.7 shows, participants who received the IAP version made significantly more upward-downwards (48.9%) movements compared to the participants who received PO (43.5%).
Table 6.7: Number and direction of saccades in a trial (hypothesis three), and difference between risk formats

<table>
<thead>
<tr>
<th>Risk format</th>
<th>Number of saccades</th>
<th>Percentage of saccades moving vertically</th>
<th>Percentage of saccades moving horizontally</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO % [α]</td>
<td>47.955</td>
<td>43.5%</td>
<td>56.5%</td>
</tr>
<tr>
<td>(2.24)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td></td>
</tr>
<tr>
<td>IAP †</td>
<td>48.228</td>
<td>48.9%</td>
<td>51.1%</td>
</tr>
<tr>
<td>Icon array difference [β]</td>
<td>0.273</td>
<td>5.4%*</td>
<td>-5.4%*</td>
</tr>
<tr>
<td>(3.17)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>440</td>
<td>440</td>
<td>440</td>
</tr>
</tbody>
</table>

Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001

6.6.6 Pupil dilation

Hypothesis and research question four sought to understand if, and how, the cognitive burden of the task differed by risk communication format through an examination of pupil size. The average (mean) pupil dilation for the overall task and to individual attributes are presented in Table 6.8. The number of observations reflects the fact that some attributes had complete visual ANA (as shown in Table 6.6). For all attributes, the mean pupil size was smaller for respondents completing the icon array version of the DCE, this was also the case for the change in pupil size for the detection and cost attributes. However, none of these differences were statistically significant.
Table 6.8: Measurements (in pixel units) of cognitive pupillometry (hypothesis four)

<table>
<thead>
<tr>
<th>Risk format</th>
<th>All</th>
<th>Probability of detecting a cancer</th>
<th>Risk of unnecessary follow-up</th>
<th>Lifetime cost of screening programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean pupil size per trial over all fixations to the choice set and fixations to each attribute, and differences between risk formats</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO % [α]</td>
<td>1228.90</td>
<td>1248.89</td>
<td>1186.24</td>
<td>1343.14</td>
</tr>
<tr>
<td>(92.24)</td>
<td>(94.44)</td>
<td>(90.46)</td>
<td>(99.75)</td>
<td></td>
</tr>
<tr>
<td>IAP ↑</td>
<td>1184.18</td>
<td>1195.84</td>
<td>1157.78</td>
<td>1264.57</td>
</tr>
<tr>
<td>Icon array difference [β]</td>
<td>-44.72</td>
<td>-53.05</td>
<td>-28.46</td>
<td>-78.57</td>
</tr>
<tr>
<td>(115.55)</td>
<td>(118.59)</td>
<td>(115.63)</td>
<td>(126.33)</td>
<td></td>
</tr>
</tbody>
</table>

Change in pupil size per trial over all fixations and fixations to each attribute, and differences between risk formats

<table>
<thead>
<tr>
<th>Risk format</th>
<th>All</th>
<th>Probability of detecting a cancer</th>
<th>Risk of unnecessary follow-up</th>
<th>Lifetime cost of screening programme</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO % [α]</td>
<td>390.33</td>
<td>216.87</td>
<td>207.69</td>
<td>130.30</td>
</tr>
<tr>
<td>(33.14)</td>
<td>(21.30)</td>
<td>(25.20)</td>
<td>(19.23)</td>
<td></td>
</tr>
<tr>
<td>IAP ↑</td>
<td>375.93</td>
<td>203.9</td>
<td>210.35</td>
<td>108.41</td>
</tr>
<tr>
<td>Icon array difference [β]</td>
<td>-14.40</td>
<td>-12.97</td>
<td>2.66</td>
<td>-21.89</td>
</tr>
<tr>
<td>(45.05)</td>
<td>(28.16)</td>
<td>(33.10)</td>
<td>(27.61)</td>
<td></td>
</tr>
</tbody>
</table>

Observations 440 439 432 376

Standard errors in parentheses, * p<0.05, ** p<0.01, *** p<0.001

6.7 Discussion

The aim of this study was to explore eye-tracking as a method to assess DCE respondents’ choice strategies. The eye-tracking data generated contributed to the results of this thesis by further exploring the difference risk format can have on the decisions respondents make in a DCE. The study is an original contribution to the literature and marks one of the first investigations into using eye-tracking as a method alongside healthcare DCEs. The next section, 6.7.1, discusses the key findings of the study before considering its strengths and limitations in section 6.7.2.

6.7.1 Key findings

The eye-tracking experiment appeared to be well calibrated with very few fixations to areas of white space. This could also suggest that the respondents were engaged in the task. The equipment was also able to record eye-tracking data for 40 of the 42 women who were invited to the experiment, a success rate of 95%.

Data on the number of fixations indicated formatting risk as either a percentage and icon array or a percentage only had no effect on the visual attention to these attributes. This indicated that there was no difference in information processing, and therefore the ability to make an informed choice. However, the attributes ‘risk of unnecessary follow-up’ and ‘probability of detecting a cancer’ both attracted more attention than the cost attribute. This
could suggest that the attributes ‘risk of unnecessary follow-up’ and ‘probability of detecting a cancer’ required more information processing or that the information was more important for the study participants. When risk was presented as an icon array rather than as a percentage only, the time spent attending to attributes was greater. This could mean participants who received IAP made a more informed choice, as information processing occurred for longer. However, a larger sample size (approximately 50 women to each risk format given a power of 0.8 and 95% confidence level) would be required to investigate the significance of this finding further.

The ANA results presented in this study matched the results found in Balcombe et al., (2014) who state: “most respondents visually attend most attributes most of the time” (p.447). Self-reported ANA was reflected in the visual attention to the ‘risk of unnecessary follow-up’ and ‘probability of detecting a cancer’ attributes. When a respondent reported that the ‘risk of unnecessary follow-up’ and ‘probability of detecting a cancer’ was not important in their choice making, they paid significantly less attention to these attributes (as measured by dwell-time).

There was no difference in dwell time between participants who reported attending the cost attribute and those who did not. The lack of a link between self-reported ANA and ANA suggested by the eye-tracking data for the cost attribute could be because respondents could have subconsciously taken account of the cost information in their choice making. Alternatively, the final questions were completed in the presence of a researcher, which could have induced a social desirability bias. The presence of this bias was also supported by the insignificant coefficient on the cost attribute in the modelling of the choice data from the eye-tracking participants, in contrast to the results of the large internet panel in Chapter Four where cost was a significant attribute in all latent preference-classes. The results of this study suggest that self-reported ANA, as measured through supplementary questions to the DCE, may be unreliable for cost attributes.

Results from the saccade data indicated that communicating risk using an icon array significantly increased the proportion of vertical eye-movements made by an individual. These differences in movements indicate that the deliberation process which led to a choice was clearly affected by the risk format. In a simple choice set experiment, with one risk attribute and one cost attribute, Arieli et al. (2011) concluded that a high proportion of vertical eye movements was indicative of an expected pay-off calculation and a high proportion of horizontal eye movements would indicate that the risk and cost were considered separately. Hypothesis three made no a priori expectations about the direction of saccades in a multi-attribute choice experiment and there was no existing hypothesis.
about the effect icon arrays would have on the movement. The results showed that whilst respondents in both risk versions of the DCEs exhibited more vertical movements than horizontal, the IAP version spent more time considering each alternative separately to make their choice.

More vertical movements could indicate that the respondent was weighing up each alternative as a whole in line with Lancaster’s Theory (Lancaster, 1966), and decision-making in line with EUT calculations. The implication of these combined economic theories is complicated (see section 2.5.4.1 of Chapter Two). As this study provided a preliminary investigation into ascertaining whether risk format affected decision strategies, there is only scope to state that a significantly different pattern occurred before a choice was made. However, the nature of the drivers of the difference in saccade patterns can only be speculated and requires further research (for example, changing the orders of the attributes, introducing non-risk attributes).

The results of section 6.5.6, indicated that pupil dilation over the whole trial was smaller for the IAP version of the DCE. This could mean that the task was less cognitively burdensome and easier for respondents to complete. However, the result was insignificant and a larger sample size would be required to investigate this finding further.

**6.7.2 Strengths and limitations**

This study used sophisticated eye-tracking equipment which recorded many variables which allows for analysis of more variables than conducted in existing published studies (Balcombe et al., 2014; Bialkova & van Trijp, 2011). However, the study had some key limitations.

Within eye-tracking, there is a trade-off between using precise equipment and apparatus that is comfortable for participants. The EyeLink® 1000 required participants to keep their head fixed and involved a forehead and chin-rest set-up. This provided an unnatural reading position and could have reminded the respondents that their eye-movements were being monitored, causing them to pay more attention to the screen than they would have otherwise. However, it is unlikely that the effects of this would differ between respondents randomised to the different risk formats.

The drift correction procedure implemented between trials to check calibration generated precise records of gaze. However, the correction procedure may have induced a starting point bias in the middle of the screen. As all participants in each risk version completed the between trial correction, the effects of risk communication format on eye-tracking outcomes should have been unaffected.
The sample size of twenty women in each condition could ideally have been larger and may possibly led to more statistical significance for some findings. With a larger sample size, covariates could have been added to the regression specifications to investigate how individual effects can influence eye-movement data. However, as described in section 6.5.1, both time and resource constraints meant that a larger sample size was not possible.

When making multiple comparisons between the two risk communication formats using the eye-tracking data, there was a risk that a difference may have occurred in their fixations, saccades or pupil dilation, by chance. The more tests conducted, the bigger the risk of this happening. This means there is an increased risk of a type I error, an incorrect rejection of the null hypothesis that there is no difference between the formats. When comparing risk communication formats, this experiment had only one statistically significant result, difference in the direction of saccades. The Bonferroni correction (Holm, 1979) for multiple comparisons adjusts the significance level to $\alpha \frac{n}{9}$, where $n$ is the number of tests. In this experiment which involved nine comparisons between the two risk communication formats, the corrected critical p-value of 0.005 ($\alpha \frac{9}{9}$) means this finding is no longer statistically significant. It is important for researchers wishing to use eye-tracking methods in the future to consider the problems associated with multiple comparisons and correct accordingly with techniques such as the Bonferroni correction or more sophisticated approaches.

A conservative limit of 75 milliseconds for a fixation was implemented as a threshold for the analysis. Other studies have used algorithms to calculate the optimal fixation threshold (Salvucci & Goldberg, 2000), but this was not conducted for this study. Similarly, sensitivity analysis using different thresholds could be used to test for robustness in the results.

In addition, randomisation of the order of choice sets or attribute location could have tested the robustness of some of the findings further. The direction of saccades could have been investigated through randomisation of the order of attributes and choice set appearance, and this would allow further investigation of a top-bottom, left-right bias.

As pupil size is also dependent on the location of the stimuli on the screen and luminosity in the lab, the cognitive burden of studies cannot be compared on either an absolute or relative scale using this metric alone. As there are various causes of pupillary responses, care must be taken to distinguish the exact activator of this response. Similarly, the luminance of the lab was not recorded for every participant and additional equipment would have allowed for a more precise measure of pupil dilation. However, the room was
windowless and all light sources were consistent between and within experiments, which would suggest a constant level of luminescence in the environment. Given the experiments were conducted under the same settings for all research participants and the two DCE stimuli occurred in exactly the same location on the screen and within the same pixel space, it is valid to make comparisons using pupil measurements in this experiment.

To some extent, participants had control over where they look at and, in theory, could game the situation by looking at particular attributes more or less in the presence of a researcher. However, pupil dilation is a measurement that respondents are unaware of and would find difficult to control. It therefore provided one of the most objective measures of mental intensity available from eye-tracking data (Laeng et al., 2012).

Extraneous variables, such as the time of day and the participants’ previous activities were not collected or controlled for in the experiment. Studies have shown that alcohol, a lack of sleep and reading may affect participants’ eye-movements. Future work, in a larger study, could collect this information from study participants. However, as respondents were randomised to different risk formats, these are unlikely to have affected the results.

Ideally, concurrent think-aloud analysis would have been conducted alongside the eye-tracking study to directly compare the two methods. Using an eye-tracking system which allowed freedom of movement and talking would have resulted in a less accurate account of visual attention. As technology progresses, future research may be able to simultaneously record eye-movements in a think-aloud interview. Alternatively, retrospective think-aloud data can be coded in line with AOI making up the themes to compare participants’ self-reported accounts and their attention (Guan et al., 2006).

**6.8 Conclusion**

This chapter presented an exploratory eye-tracking study which looked at the effect risk communication format had on DCE respondents’ choice strategies. The study was innovative in its design, as very little previous research has been conducted using eye-tracking alongside the completion of a DCE. This study found that eye-tracking offered a promising method to understand more about how DCE respondents viewed the choice task. The study also found that visual attention to attributes was greater, and cognitive burden lower, when risk was presented with an icon array; however these differences were not statistically significant (probably due to sample size). Self-reported ANA was only reliable for the non-cost attributes, suggesting respondents may not be truthful about their attention to these attributes. There were statistically significant differences in how participants searched for information in the choice set to make their decision between risk formats.
However, research combining eye-tracking and choice surveys is very much in its infancy, and thus there is potential to answer many further research questions and other risk format effects in future work.
Chapter Seven
Discussion

7.1 Introduction
The thesis examined how the communication of risk attributes in a DCE eliciting women’s preference for breast screening affected their choices and decision-making heuristics. Systematic review methods, discrete choice modelling, qualitative research methods and eye-tracking methods were used to investigate how DCEs with risk attributes were completed by respondents and the subsequent values derived. The use of mixed research methods aimed to provide a comprehensive approach to meeting the overarching aim; how to communicate risk in a healthcare DCE.

In section 7.2, the key findings from each chapter are summarised and discussed with regards to their relative contributions. The limitations of the thesis are described in section 7.3, with a discussion of the lessons for future research and researchers using DCEs. In section 7.4, the results are contrasted with relevant existing literature. The key implications and generalisable findings which may be useful for other researchers or policy-makers are presented in section 7.5. Options for future research which were beyond the scope of this thesis are described in section 7.6. Finally, in section 7.7, the chapter reflects on the contribution of this thesis.

7.2 Key results
The first research question concerned the extent to which qualitative research was reported in healthcare DCEs and how useful these research methods were when used alongside DCEs. The systematic review of qualitative research methods alongside healthcare DCEs, and the survey to authors of these studies, identified qualitative research as an under-reported yet useful method. Out of the 254 healthcare DCEs identified, only 29 studies reported the details of the qualitative component in-depth. There was no qualitative research reported in 111 of the identified studies. Despite the lack of reporting, the findings of the survey to authors indicated that all authors believed that using qualitative research methods added value to their study. For identifying attributes and levels, the most common approach were interviews or focus groups. For understanding more about thought processes, decision strategies and respondents’ understanding of the task, the ‘think-aloud’ method was identified as most useful. In reviewing the quality of qualitative research conducted alongside DCEs in the 29 studies which reported details, it became apparent that conducting qualitative research alongside stated preference methods was a unique, untraditional, application. As a result, existing appraisal tools for qualitative research...
methods were found not to be applicable in this setting. A bespoke appraisal tool was therefore developed in order to extract this information consistently and in a systematic, but fairer, manner than existing tools.

The second research question sought to identify the most effective risk communication formats which may aid respondent decision-making in a DCE. To answer this research question, Chapter Three (section 3.2) and Appendix 3.1 presented a structured review of risk communication formats. The review identified icon arrays as the most promising risk communication format which was found to be most effective, compared to other formats, in 10 of the 29 empirical studies which investigated its use as a risk communication tool. However, the structured review also found that the evidence supporting the potential formats was highly heterogeneous. Understanding the relative merits of the different formats, the populations in which they could be most effective, and the best criteria for assessing efficacy is a point for further research. For example, investigations using randomised designs directly comparing formats with appropriate outcomes for appraising the efficacy would be useful. This chapter also identified other challenges in communicating risk information and assessing the effectiveness of the communication format. The structured review of formats generated some incidental findings with regards to the predictors of individuals’ ability to interpret risk information, namely: people’s perceived risk; their experience of the risk or hazard; and their numeracy skills.

The third research question sought to understand if, and how, the risk communication format of attributes in a DCE affected the valuations derived. In order to investigate this research question, a case study was selected: women’s preferences for breast screening. In Chapter Four, a large internet panel was surveyed to elicit female members of the public’s preferences. Women were randomised to two risk communication formats: 1) percentages only (the most commonly used format in healthcare DCEs); and 2) icon arrays and percentages. Estimation of a series of random utility choice models indicated that the risk communication format did not affect either women’s marginal utility or choice consistency. However, women’s preferences were highly heterogeneous (some women most sensitive to cost; others to the probability of detecting a cancer; and others to the risk of unnecessary follow-up) with six latent preference-classes identified. Significant predictors of class membership were women’s concern about their own risk of developing breast cancer and their ethnicities. The finding that the valuations of breast screening attributes derived from this DCE were insensitive to the risk communication format was surprising given the findings of the structured review presented Appendix 3.1. However, a
key implication was the importance of understanding DCE respondents’ attitude towards
the risk associated with the good, service or intervention being valued.

Chapter Five utilised qualitative research methods as an alternative approach to traditional
quantitative choice modelling, to investigate the sensitivity of DCE respondents’ choices
and decision strategies to the risk communication method in the DCE. Qualitative research
methods have not previously been used to investigate attribute format effects in DCEs. The
think-aloud interviews revealed that when risk was communicated as a percentage only
women often visualised the risk, creating their own icon array and relieving part of the
task’s burden. There were often sentiments of initial ‘panic’ when the percentages only
version was completed by the respondents, and the icon array was much better received.

The fourth research question aimed to understand how effective eye-tracking methods
were in evaluating respondents’ understanding of DCE attributes and levels. Chapter Six
used eye-tracking methods to understand more about individuals’ decision strategies and
whether risk communication format affected these. The eye-tracking experiment possessed
characteristics of validity, with successful calibration and little attention to areas of no
information. The eye-tracking results revealed significant differences between the DCE
versions, with women receiving icon arrays making significantly more up-down eye-
movements when they completed the choice task. The consequence of this finding on the
underlying economic theories, which support the use of DCEs as a valuation method,
requires further research.

Results for other outcomes suggested that women receiving icon arrays also paid more
visual attention to attributes (indicating more information processing) and found the DCE
less cognitively burdensome. However, differences between risk formats were not
statistically significant for these outcomes. Interestingly, although women who self-
reported non-attendance to the risk and detection attributes visually attended these
attributes significantly less than those who reported attending them, this result was not
found for the cost attribute. The mismatch of actual and self-reported attribute non-
attendance to the cost attribute (and not in other attributes) could indicate some bias due to
researcher presence, such as social desirability. In this respect, eye-tracking data could
prove more reliable than self-reported behaviour.

Each of the three empirical studies contributed to the overall aim which was to understand
if and how DCE respondents’ choices and decision strategies were affected by the risk
communication format. In the case of understanding whether choice was affected, the
insignificant parameter on the scale term in the heteroscedastic conditional logit and the
finding that communication format did not predict scale or preference class membership suggests that the addition of icon arrays made no difference. Alternative research methods (qualitative and eye-tracking) confirmed some of the findings of the quantitative study but also suggested that icon arrays relieved some of the cognitive burden in the choice task.

7.3 Limitations
In this section, the limitations of each study conducted in this thesis will be discussed. Possible remedies to the issues will also be discussed where possible.

The systematic review of qualitative research conducted alongside DCEs which was presented in Chapter Five (section 5.2) and Appendix 5.1 used standardised systematic review methods and published guidelines. Two potential limitations were identified: 1) the reliance on material published in the articles; and 2) the lack of an appropriate appraisal tool. These limitations were rectified with a survey to authors and the creation of a bespoke appraisal tool. However, the survey to authors only involved contacting authors who had reported basic qualitative research in their studies and did not include those who did not report any. Authors who did not report qualitative research in their study frequently mentioned the absence of qualitative methods as a limitation, suggesting they felt it would have added value. Authors who reported extensive details about the qualitative aspect were assumed to have presented the research as it was an important component of their study. In hindsight, it may have been useful to extend the survey to all authors of the review (particularly to understand the views of authors who reported no details). As a consequence, the survey results may be biased and authors who did not report qualitative details (and hence were not contacted) might have done so because this element was not important or useful.

Another limitation of the systematic review presented in section 5.2 and Appendix 5.1 was that the data extraction was also only conducted by one author (CV). In retrospect, it would have been useful to have a second reviewer extracting data. However, the recent publication of a systematic review employing the same strategy (Clark et al., 2014) in part validates the quality of the systematic review presented in Chapter Three, which identified more DCE studies.

In Chapter Three (section 3.2), the review used a specific search strategy to rapidly identify a range of risk communication formats. The advantage of using this approach was that the review was conducted promptly. However, this meant that non-health risk communication was not covered. For example, future research could explore environmental risk communication or literature in health and safety, or engineering. The analysis of the
extracted data was also limited. Counting the best performing risk communication formats resulted in more frequently used formats appearing more popular. In hindsight, a more sophisticated meta-analysis of each empirical study could have been conducted with a thorough ordering of the formats rather than a best-worst rating. It is not clear if, or how, using this more sophisticated approach would have changed the results of the review.

In the large sample DCE study described in Chapter Four, an internet panel provider was chosen to yield responses. Internet panels were chosen to acquire a large sample size, paramount for the investigation of heterogeneity, quickly and relatively inexpensively. However, there are limitations to this sampling approach. The respondents to internet panel DCEs are likely to be computer-literate and therefore could be biased against the older-age ranges. Furthermore, the respondents were paid to complete the survey which could have effected their engagement in the task and causing them to provide ‘answers’ to things they would not chose in real life. There is evidence to suggest internet panels condition respondents whose preference and behaviour may change due to participation in the panel (Dennis, 2001). The advantages and disadvantages of using internet panels for DCEs have yet to be explored; however, other health valuation studies have found they provide good quality data compared to other methods such as postal surveys, telephone interviews (Mulhern et al., 2015). It is, however, acknowledged that the views found in this study may not necessarily be representative of the general public’s which may consequently hinder the interpretation of the quantified benefit-risk trade-offs and even the existence of the distinct latent preference-classes identified.

Although other discrete choice models could have been fitted (such as the MXL or nested logit), the study used sophisticated econometric techniques beyond that which are typically used in healthcare DCEs (Clark et al., 2014). It was believed that describing heterogeneity in terms of latent class models would be most useful for decision or policy-makers. In addition to providing details about preference heterogeneity across groups, rather than across all individuals, latent class models have also outperformed mixed logit models in most tests of statistical properties (Shen, 2009).

The qualitative interviews presented in Chapter Five used a verbal-protocol method which was particularly susceptible to three key issues: reactivity, nonveridicality and reflexivity. The study presented in Chapter Five recognised the impact of the interviewer’s presence on the interviewees’ choices, and accounts of making choices, and hence the analysis and interpretation of the qualitative data collected. Efforts were made to minimise the researcher bias with the flexible interview schedule which largely comprised prompts and the deductive analytical approach of open-coding. The framework to hone themes into the
categories of ‘risk’ and ‘strategy’ were to focus the analysis but did not influence the emergence of new themes. However, it is impossible to say that, even an unconscious, awareness of the background research question did not influence the reading of transcripts.

Women who participated in the qualitative interviews made contact in response to an advertisement (email, website, or noticeboard). It could therefore be that the sample in this study was biased toward middle-class and those who were intrinsically motivated. This problem is not unique to this study; the challenges of recruiting diverse research participants are well established, particularly for studies with a health focus (Yancey et al., 2006; Shavers-Hornaday et al., 1997).

The eye-tracking study presented in Chapter Six suffered from a small sample size chosen based on previous studies (Bialkova & van Trijp, 2011), of which there were very few. Retrospective sample size calculations suggest that approximately 100 women would be required to identify statistically significant differences in outcomes between the risk communication formats. The sample size in this study was also restricted by time (due to the shared laboratory) and resource constraints. The recruitment of eye-tracking participants was particularly challenging as women had to visit the university’s campus during core weekday hours. This was unavoidable due to the laboratory set-up and sharing arrangements. As a consequence, women who participated in this study may not be representative on the general public. This is acknowledged as a limitation of this study. More time and resources could have resulted in an increase in the sample size and therefore an increase in the meaningfulness of the results through statistical significance.

The device used in the eye-tracking study required respondents to rest their chin and forehead for stability. This unnatural set-up may have interfered with the respondents’ behaviour or choices. However, less intrusive devices, recording with a lower frequency, would have compromised the accuracy of the eye-tracking data. Furthermore, many remote devices do not record data on pupil size, meaning an examination of the effect of risk communication format on cognitive burden could not have been conducted.

Comparisons between the results of the three empirical studies of this thesis was challenging due to the differences in the epistemological and ontological assumptions under which the research was conducted. Similarly, comparing conflicting findings requires an opinion of which method was most valid. In health services research, there is a lack of consensus even with regards to the synthesis of mixed-methods findings (Brannen, 2006). As a consequence, the thesis is restricted in its ability to compare findings across the empirical chapters.
All empirical studies employed the same DCE containing just three attributes. In a more complex experiment, the effect of icon arrays on respondents’ marginal utilities and decision-making heuristics may have been different. Other attributes, such as process attributes, could have been included; however, these three attributes were chosen based on literature reviews and discussions with experts in breast cancer. Including more attributes may have increased the cognitive burden for respondents and detracted from the risk attributes. A simple DCE also resulted in large areas of interest which enabled a clear analysis of the eye-tracking data. Similarly, choosing a familiar example was a key criterion for selecting a case study in Chapter Three in order to limit any other cognitive burden to the respondents.

The use of round numbers for the risk attribute (1%; 5%; 10%; and 20%) are arguably easy for women to visualise even when presented only in percentage form. Therefore a greater test of the benefits of presenting an additional icon array may have been investigated through smaller percentages (<1%) or presenting them to the nearest one thousandth (for example, X.X%). Furthermore, the empirical studies in this thesis only explored two risk communication formats out of the numerous pictorial, graphical, numerical and qualitative approaches identified in the structured review of section 3.2, Chapter Three.

It is acknowledged that the event or hazard described can affect the interpretation of its likelihood. It has been found that a 1 in 100 risk of death feels more likely than a 1 in 100 chance of a cold even if, statistically, they are the same (Sunstein, 2003). This is known as ‘probability neglect’ and highlights the importance of the event on people’s interpretation of its likelihood (Zeckhauer & Sunstein, 2009). With a different event and even the same probabilities, the results of this thesis could be very different. For example, if the risk of unnecessary follow-up was changed to risk of mortality (keeping the probabilities the same), the effect of the risk communication format may have been more prominent.

Stated preference methods are limited in that they rely on respondents saying what they would do in a hypothetical situation. However, in a publically-funded healthcare system there are rarely market situations from which revealed preference data can be generated. Therefore, a robust and established stated preference method was chosen using a choice-based format. Novel, less established, methods may have induced other limitations. The hypothetical nature of stated preference valuation methods and new alternative choice experiment designs, such as BWS, are recognised as a limitation of this thesis.

Furthermore, all stated preference methods involve the use of surveys. The empirical studies of this thesis relied on a survey based method which therefore implicitly resulted in
a literate sample. The different mode of administration (web-based, face-to-face and eye-tracking) and the different choice results (for example cost becoming insignificant in a face-to-face setting) suggest that surveys are affected by their administration. This is in line with other survey-related literature (see Hyman (1944); Preisendörfer & Wolter (2014)). As a survey design is inherent to stated-preference studies, and revealed preference data was not available, the issues arising from a survey-based approach in this context are unavoidable.

7.4 Comparisons with existing literature
There was little existing research investigating the communication of risk attributes in DCEs. The results of other studies that have only explored quantitatively the effects of risk communication on choices (Howard & Salkeld, 2009; Buchanan et al., 2014), had major limitations in that they failed to account for differences in the scale parameter when comparing the risk communication formats. The analysis conducted in Chapter Four extensively accounts for differences in scale, and explores preference heterogeneity, another area also under-investigated in existing literature (Clark et al., 2014).

The risk communication literature reviewed in Chapter Three (section 3.2) was supportive of the use of icon arrays as a risk communication format. The results of this study found that icon arrays were well received by DCE participants, even if they did not directly affect the valuations. Other components of important aspects of risk communication identified in the literature were important. The quantitative study, in particular, highlighted the importance of women’s perception (or concern) on their risk preferences.

Literature on women’s preferences for screening has suggested that women are tolerant of the risk of unnecessary follow-up, even at rates of 30% (Baum, 2013). The quantification of preferences in this thesis showed that although some women were very tolerant, others were not. The heterogeneity in preferences match other literature which suggests large variation in people’s preferences for healthcare (Hole, 2008).

7.5 Key implications
The culmination of the findings from each of the studies in this thesis provides important implications for researchers and policy-makers. These implications are now described.

For researchers seeking to use DCEs for eliciting individuals’ preferences for benefit-risk trade-offs, the method was not sensitive to the risk communication formats used in this study. The qualitative study identified that respondents felt more engaged with the task when risk was presented with an additional icon array. As a result, this PhD suggests that
icon arrays in addition to percentages are preferred to percentages alone when presenting risk attributes in a DCE.

Another key implication for researchers was the usefulness of qualitative research methods alongside DCEs. The survey to authors found all researchers’ experience of the methods added value to their DCE. Given the lack of qualitative research conducted alongside DCES, previous calls for qualitative research in health economics (Coast, 1999) and alongside DCEs (Coast et al., 2012) appear to have remain unanswered. The bespoke appraisal tool developed as part of this thesis and presented in Appendix 5.5 may be a useful checklist for health economists seeking to conduct qualitative research alongside their stated preference study. The appraisal tool may also be incorporated into existing guidelines, such as those published by ISPOR, or utilised by researchers in the form of standardised reporting criteria.

In this thesis, eye-tracking was successfully used alongside a DCE and was found to be a valid and insightful method for examining respondents’ decision-making heuristics. A recommendation of this thesis is that eye-tracking should be used in future research to understand more about choice-making behaviour and attention to attributes. Researchers seeking to use eye-tracking methods alongside DCEs can do so, but should be aware of the large sample sizes required to achieve significance and the relative advantages and disadvantages of different types of eye-tracking devices.

Women’s preferences for an NHSBSP were highly heterogeneous. For researchers, models which allow for such preference heterogeneity need to be used when modelling DCE choice data. Particularly researchers in health, where applications of these methods are limited (Clark et al., 2014). Additionally, the scale parameter must also be considered if there are reasons to believe respondent characteristics might influence choice consistency. Failure to account for differences in scale may prevent accurate investigation of preference heterogeneity. Therefore researchers modelling healthcare choice data should consider greater use of scale-adjusted latent-class and heteroscedastic conditional logit models.

There are also implications for policy-makers wanting to implement an effective NHSBSP. Policy-makers must consider that there is a substantial group of women who were less tolerant of the risks of screening, and recognise that ethnic minority groups place little value on the benefit of detecting a cancer. If policy-makers wish to improve up-take in minority groups, then the benefits of screening need to be better communicated or incentives should be in place to encourage programme participation. However, ‘informed
choice’ is a rising consideration for policy-makers providing healthcare services (Forbes & Ramirez, 2014); and increased participation may not be an important objective.

Policy-makers should also realise that some women received a very high utility from just participating in a NHSBSP, regardless of the benefits, risks and costs. This finding suggests women placed a high value on non-health benefits. Decision-makers may wish to consider alternative valuation methods to extend or complement the results of a traditional QALY based cost-utility or cost-effectiveness analysis, to incorporate alternative sources of utility. As NICE currently makes decisions using an extra-welfarist view-point and the results of CUA and/or CEA, the NHSBSP may be undervalued under this framework.

Recent changes to breast screening policy have included an extension of the screening ages by six years, inviting women between 47 to 73 years old in some areas of England (NIHR Trial ISRCTN33292440). Policy-makers should be aware of women’s preferences for the increased risk of unnecessary follow-up when screening younger and older women. For women between 50 and 70 years, how preferences change through the programme (as the probability of detecting a cancer reduces and the risk of unnecessary follow-up rises) also needs to be considered by policy-makers (Barratt, 2015).

7.6 Future research
The studies presented in this thesis can be taken forward in several ways.

The critical appraisal tool developed for assessing the quality of the qualitative research reported alongside published DCEs should be trialled with different stated preference studies such as BWS and with more contingent valuation studies. Successful application of this tool in more research areas would provide scope to transform the tool into best practice or reporting guidelines.

How the analysis of DCE choices can incorporate other expected utility theories is still under development and requires further research. EUT can be combined with random utility theory which models the uncertainty in conclusions drawn about the unobservable error component of utility (Polak & Liu, 2006). Exploration of this in a healthcare setting has identified this as an important consideration (Robinson et al., 2015). In addition, alternative theories for analysing choice data, such as regret minimisation, may be appropriate (Thiene et al., 2012).

The effect of different risk frames for attributes in a DCE (for example, survival versus mortality) has been investigated previously (Howard & Salkeld, 2009). Given the existence of research already investigating framing-effects, this thesis concentrated on the
effect of risk communication formats specifically. Future research could consider the effect of risk communication format and risk frame, for example randomising respondents within each format to receive the ‘probability of receiving a true positive’ (in contrast to attribute ‘probability of unnecessary follow-up’).

Not only does the risk attribute need to be appropriately communicated, but for WTP calculations, cost should be communicated appropriately too. The think-aloud method revealed other strategic behaviour (such as using others’ budget constraints) and the presence of non-use values in women’s decision-making indicating that current communication methods may be improved. Further investigation of the presence of these biases and the appropriate method of communicating cost in a healthcare setting is warranted. This could be conducted via think-aloud methods or other qualitative approaches to ascertain the presence of such phenomena.

The use of eye-tracking as a method alongside DCEs was very exploratory and there is scope for further research in this area. Future studies should consider using a larger sample size to investigate visual attention to information in a DCE choice set. Other research could extend the use of pupil dilation as a proxy for task complexity in a DCE setting by comparing pupil size in choice-sets containing alternatives of similar utility (a difficult choice) where correct communication is likely to be most important. Further research may involve the incorporation of visual attention data into the quantitative modelling of choices, or a comparison of eye-tracking attention to qualitative accounts of behaviour collected in retrospective think-aloud interviews.

7.7 Conclusion
This thesis presents a contribution to multiple research areas including: the applied DCE literature; eye-tracking and psychology research; and literature utilising qualitative research methods alongside stated preference methods. Whilst few definitive answers can be given, an insight into the usefulness of different research methods and the effect of risk communication format on choices and choice-making in a DCE context has been extensively tested. It appears that DCEs can be used to value benefit-risk trade-offs. These benefit-risk trade-offs were not sensitive to the risk communication format of the attributes in the selected case study. However, providing an icon array is highly recommended with the results of the qualitative study suggesting this is beneficial for DCE respondents. In addition, the results and implications of this thesis are likely to be of interest to both researchers and policy-makers.


Kennedy, A., Gately, C. and Rogers, A. (2005). Examination of the implementation of the Expert Patients Programme within the structures and locality contexts of the NHS in England (PREPP study) - Repoer II. *National Primary Care Research and Development Centre*.


Microsoft (2010). *Excel*.


Sawtooth (2012). *Sawtooth Software SSI Web 8.3.8.*


SR Research (2013). Data Viewer Software.

SR Research (2012). Experiment Builder Software.

StataCorp (2011). Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.


Using discrete choice experiments to value benefits and risks in primary care

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Medical and Human Sciences

2015

Caroline Mary Vass
School of Medicine

Volume II of II
Volume II

List of Appendices
Appendix 3.1: How to frame risk in a DCE: an exploration and structured review ........246
Appendix 3.2: List of studies included in the structured review of risk communication formats ..............................................................................................................................................268
Appendix 3.3: The published paper containing the results of the structured review ........274
Appendix 3.4: List of DCEs with a focus on breast cancer and screening from the systematic review ..........................................................................................................................................................................................295
Appendix 3.5: The design of the DCE used in the pilot study ........................................297
Appendix 3.6: Review of questions used to elicit individual characteristics in commonly-used surveys ........................................................................................................................................................................................................300
Appendix 3.7: Confirmation letter of University Research Ethics approval ..................302
Appendix 3.8: The DCE survey used for the pilot study .............................................303
Appendix 3.9: Online advertisement for study participants ........................................339
Appendix 3.10: Preliminary interview schedule for pilot study ..................................340
Appendix 3.11: Information sheet for interviewees for the qualitative interview study ...341
Appendix 3.12: Consent form for interviewees for the qualitative interview study ..........343
Appendix 3.13: NHS Choices video approval ................................................................344
Appendix 3.14: Pilot quantitative results of the DCE from the internet panel ................345
Appendix 3.15: The design of the DCE used in the final empirical studies ...................347
Appendix 3.16: The DCE survey used for the final empirical studies ............................349
Appendix 3.17: Quantitative results of the DCE from the internet panel in the second pilot study ..................................................................................................................................................................................................................................................387
Appendix 4.1: ResearchNow® response rate in the large internet panel DCE ..............388
Appendix 4.2: Sample characteristics of respondents who completed the DCE in the final study ..................................................................................................................................................................................................................................................389
Appendix 4.3: Results of a simple logit regression to determine if any covariates were a significant predictor of risk communication format ..........................................................................................................................................................................................392
Appendix 4.4: Latent class analysis results from 6-preference-class (no scale-classes) model for the large internet panel DCE ..................................................................................................................................................................................................................................................394
Appendix 5.1: Systematic review of qualitative research alongside healthcare DCEs ....395
Appendix 5.2: Survey to authors of published healthcare DCEs ..................................420
Appendix 5.3: Flow of papers through the systematic review ........................................428
Appendix 5.4: Comparison of existing appraisal tools for assessing the quality of qualitative research

Appendix 5.5: Final critical appraisal tool for assessing the quality of qualitative research alongside DCEs

Appendix 5.6: Studies included in the systematic review of qualitative research conducted alongside DCEs

Appendix 5.7: Tabulated results from the systematic review of DCEs

Appendix 5.8: List of studies not included in the Clark et al. (2014) systematic review

Appendix 5.9: The submitted paper describing the development of the appraisal tool

Appendix 5.10: The questionnaire sent to authors of DCE studies to assess their experience and views of using qualitative research methods

Appendix 5.11: The invitation to authors of DCE studies to complete the questionnaire to assess their experience and views of qualitative research methods alongside DCEs

Appendix 5.12: The reminder invitation to authors of DCE studies to complete the questionnaire to assess their experience and views of qualitative research methods alongside DCEs

Appendix 5.13: Author responses to the questionnaire of their experience and views of conducting qualitative research alongside DCEs

Appendix 5.14: Paper advertisement to the general public for participation in the think-aloud interviews

Appendix 5.15: Email advertisements to The University of Manchester staff for participation in the think-aloud interviews

Appendix 5.16: Think-aloud interview schedule

Appendix 5.17: Flow of recruited study participants for the think-aloud interviews and the eye-tracking study

Appendix 6.1: Paper advertisement to the general public for participation in the eye-tracking study

Appendix 6.2: Email advertisement to The University of Manchester staff for participation in the eye-tracking study

Appendix 6.3: Information sheet for the eye-tracking study participants

Appendix 6.4: A description of the eye-tracking pilot study

Appendix 6.5: Consent form for participants in the eye-tracking study

Appendix 6.6: iPad survey for the eye-tracking study

Appendix 6.7: Heteroskedastic conditional logit results for the eye-tracking study
Appendix 3.1: How to frame risk in a DCE: an exploration and structured review

Aims
The overall aim of this review was to systematically identify different formats which can be used to communicate risk information to DCE respondents. If risk is communicated effectively, then it will assist the respondents in making informed choices based on their preferences and thus improve the quality of the data collected. The review is described as ‘structured’ to distinguish it from the systematic review presented in Appendix 5.1, and because no second screening of abstracts was conducted.

Methods
The following section describes the steps taken to identify studies using different risk communication formats.

Search strategy
A systematic search of five electronic databases (Medline, Embase, Web of Science, PsychINFO and EconLit) was conducted in April 2013. The search checked the titles of articles in each database for the terms risk and either communication or format, and the whole abstract for topic terms (aid, presentation, display). The exact strategy included the following terms: Title=(risk*) AND Title=(communicat* or format*) AND Topic=(aid* or present* or display*).

The search strategy aimed for a low recall and very high precision and was developed based on extensive meetings with an expert librarian (personal communication with Mary Ingram). The databases were searched for studies published between January 2000 and April 2013. Any valuable communication formats reported in papers published prior to 2000 would likely be reproduced or compared in a later study which would be captured by the selected time horizon.

Screening
An initial title and abstract screen to select relevant studies was conducted. If studies could not be rejected with certainty, the whole article was retrieved and reviewed. Studies were eligible for inclusion if they: 1) tested or compared risk communication formats which aimed to aid people’s understanding of quantitative information; 2) related to a healthcare risk (the probability of a health-related outcome); and 3) were published in peer-reviewed journals written in English. Any studies which were not peer-reviewed, were not written in English were excluded. Studies reviewing public health risks, such as nationwide vaccines,
were also excluded. This was because public health campaigns often looked at communicating risk to promote a particular behaviour or change preferences (for example, describing cancer risks to encourage smoking cessation) rather than objectively providing information. Studies not using written communication, such as interactive games, were also excluded because of programming challenges meaning they are not easily incorporated into a typical DCE.

After retrieving all the full papers, the studies were initially categorised into four categories: 1) empirical studies testing risk communication formats; 2) systematic reviews of risk communication formats; 3) papers providing an overview; and 4) studies specifically communicating uncertainty around point estimates. In this review no specific assessment for methodological quality was conducted; the studies were so different that no appropriate appraisal criteria existed.

Data extraction
Empirical studies comparing different risk communication formats were viewed to be of most interest. Data from these empirical studies were tabulated using the following headings: risk communication format; outcome measures; sample size; sample characteristics; and study conclusion. However, studies were also included if they provided a review or overview of different methods. Studies which looked at communicating uncertainty rather than risk were also acknowledged but did not undergo data extraction for the reasons explained in. The identified review pieces, overview papers and uncertainty studies are drawn upon in the discussion in section of this review.

Results
Figure A1 summarises the results of the search strategy and screening process (details about the search strategy and results can be found in Table A1). The search of all databases revealed 1,207 possible hits. After the removal of non-peer-reviewed material and duplicates, 390 titles and abstracts were reviewed. In total, 215 full articles were retrieved for further review and 99 papers were included in the final review.
Table A1: Search results from all databases for the systematic review of risk communication methods, run on the 23rd April 2013

<table>
<thead>
<tr>
<th>Database</th>
<th>Search Terms</th>
<th>Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web of Science</td>
<td>Title=(risk*) AND Title=(communicat* or format*) AND Topic=(aid* or present* or display*) [limited 2000-2013; English Language]</td>
<td>482</td>
</tr>
<tr>
<td>Medline</td>
<td>risk*(ti) AND (communicat* or format*)(ti) AND (aid* or present* or display*)(ti) [limited 2000-2013; English Language]</td>
<td>223</td>
</tr>
<tr>
<td>Embase</td>
<td>risk*(ti) AND (communicat* or format*)(ti) AND (aid* or present* or display*)(ti) [limited 2000-2013; English Language]</td>
<td>313</td>
</tr>
<tr>
<td>EconLit</td>
<td>risk*(ti) AND (communicat* or format*)(ti) AND (aid* or present* or display*)(ti) [limited 2000-2013; English Language]</td>
<td>7</td>
</tr>
<tr>
<td>PscyhINFO</td>
<td>risk*(ti) AND (communicat* or format*)(ti) AND (aid* or present* or display*)(ti) [limited 2000-2013; English Language]</td>
<td>182</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>1207</td>
</tr>
<tr>
<td></td>
<td>Removal of duplicates</td>
<td>470</td>
</tr>
<tr>
<td></td>
<td>Removal of books, patents, dissertations</td>
<td>390</td>
</tr>
</tbody>
</table>

Of the studies excluded, most were either not exploring risk communication, were not looking at healthcare risks or were not using written risk communication formats.

Appendix 3.2 list all 99 studies included in the review. Each included study was given an ID comprising a letter to indicate the type of study (R for risk; U for uncertainty; O for overview papers; and S for systematic reviews) and a number for identification. Of the included studies, 65 were empirical tests of risk communication formats, 21 were overviews of the risk literature with no reported search strategy, nine were systematic reviews with a structured search strategy, and four looked at presenting uncertainty (for example, the use of confidence intervals as opposed to point estimates).
Figure A1: Study selection process
**Risk communication formats**

The 65 empirical studies revealed a number of different ways of communicating risk information. Table A2 summarises the empirical studies identified by the review with a summary of the formats compared and which formats were found to be most and least effective as reported by the study authors.

The most commonly compared formats were numerical presentations of risk statistics such as percentages, natural frequencies (NF), decimals or ratios (n=55, 85%) either alone or in combination. When risk was presented as a frequency, multiple methods were tested including formatting the probability as X in 100 and 1 in X.

Other risk communication formats tested were visual images such as matrices of coloured shapes called ‘icon arrays’ (n=29, 45%) (these can also be referred to as ‘risk grids’ or ‘pictographs’). Risk ladders were used in ten studies (15%) where the assigned risk was presented on a scale ranking it along with other probabilities (this was referred to as a risk continuum in R3; a risk thermometer in R47; and involved a magnifying graphic in R60). Two studies (R32, R36) used a risk ladder called the Paling Perspective Scale (Paling, 2003).

A number of studies also used graphical information (n=15, 23%), presenting the risk in bar charts (both horizontal and vertical), pie charts or line graphs. 11 studies (17%) used epidemiological measures to explain the risk. This included: number needed to treat (NNT) which is the inverse of the absolute risk reduction (ARR); number needed to harm (NNH) which is the inverse of the attributable risk when the NNT is negative; relative risk reduction (RRR); and survival curves (Hutton, 2009).

Thirty four studies (52 %) used qualitative descriptions of the risk, verbal analogies or vignettes. The verbal descriptors were sometimes used alone but were often combined with percentages (see R4, R8 and R52). R40, R44 and R59 used a traffic light to represent the risk information indicating low (green), medium (amber), and high (red) risk values.
Table A2: Characteristics of included empirical studies

<table>
<thead>
<tr>
<th>ID (year)</th>
<th>Type of risk communication format examined</th>
<th>Overall conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Numerical format</td>
<td>Frequency format</td>
</tr>
<tr>
<td>R1 (2002)</td>
<td>Vignette</td>
<td>Personal vignettes</td>
</tr>
<tr>
<td>R2 (2004)</td>
<td>%</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R3 (2012)</td>
<td>% and frequency</td>
<td>X in 100</td>
</tr>
<tr>
<td>R4 (2009)</td>
<td>% and frequency</td>
<td>X in 100</td>
</tr>
<tr>
<td>R5 (2009)</td>
<td>Frequency</td>
<td>NF</td>
</tr>
<tr>
<td>R6 (2010)</td>
<td>% and frequency</td>
<td>1 in XXX</td>
</tr>
<tr>
<td>R7 (2008)</td>
<td>% and frequency</td>
<td>1 in XXX and X in XXX</td>
</tr>
<tr>
<td>R8 (2007)</td>
<td>%</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R9 (2008)</td>
<td>Bar chart</td>
<td>Icon arrays, flow chart</td>
</tr>
<tr>
<td>R10 (2006)</td>
<td>Icon arrays, ladder**</td>
<td>Comparators</td>
</tr>
<tr>
<td>R11 (2004)</td>
<td>Frequency</td>
<td>X in 1000</td>
</tr>
<tr>
<td>R12 (2008)</td>
<td>% and frequency</td>
<td>1 in X</td>
</tr>
<tr>
<td>R13 (2001)</td>
<td>Bar chart, Line graph</td>
<td>Icon arrays, ladder</td>
</tr>
<tr>
<td>R14 (2008)</td>
<td>Verbal descriptors</td>
<td>%</td>
</tr>
<tr>
<td>ID</td>
<td>Type of risk communication format examined</td>
<td>Overall conclusion</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td>Numerical format</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency format</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graphical risk</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pictorial risk</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Qualitative format</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Epidemiology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Most effective</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Least effective</td>
<td></td>
</tr>
<tr>
<td>R15</td>
<td>%</td>
<td>Prolongation of life</td>
</tr>
<tr>
<td>R16</td>
<td></td>
<td>Comparators</td>
</tr>
<tr>
<td>R17</td>
<td>%</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R18</td>
<td>%</td>
<td>Bar chart</td>
</tr>
<tr>
<td>R19</td>
<td>Frequency</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R20</td>
<td>Frequency</td>
<td>X in 10,000, 1 in XXXX and XX in XXXX</td>
</tr>
<tr>
<td>R21</td>
<td>Frequency</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R22</td>
<td>Frequency</td>
<td>1 in XXX</td>
</tr>
<tr>
<td>R23</td>
<td>% and frequency</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R24</td>
<td>% and frequency</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R25</td>
<td>% and frequency</td>
<td>X in 1000</td>
</tr>
<tr>
<td>R26</td>
<td>Frequency</td>
<td>1 in XXX and X in 10,000</td>
</tr>
<tr>
<td>R27</td>
<td>Frequency</td>
<td>X in 100</td>
</tr>
<tr>
<td>R28</td>
<td>Frequency</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R29</td>
<td>%</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R30</td>
<td>%</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>ID (year)</td>
<td>Type of risk communication format examined</td>
<td>Overall conclusion</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td></td>
<td>Numerical format</td>
<td>ARR and RRR</td>
</tr>
<tr>
<td>R30 (2011)</td>
<td>Frequency</td>
<td>ARR</td>
</tr>
<tr>
<td>R31 (2012)</td>
<td>Frequency</td>
<td>RRR</td>
</tr>
<tr>
<td>R32 (2009)</td>
<td>Ratios</td>
<td>Frequency</td>
</tr>
<tr>
<td>R33 (2009)</td>
<td>Frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R34 (2008)</td>
<td>Frequency and decimal</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R35 (2009)</td>
<td>Frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R36 (2003)</td>
<td>Frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R37 (2006)</td>
<td>% and frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R38 (2012)</td>
<td>Frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R39 (2009)</td>
<td>Frequency and ratios</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R40 (2011)</td>
<td>% and frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R41 (2011)</td>
<td>% and frequency</td>
<td>%</td>
</tr>
<tr>
<td>R42 (2011)</td>
<td>Frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R43 (2013)</td>
<td>Frequency</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R44 (2009)</td>
<td>%</td>
<td>No clear conclusion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of risk communication format examined</th>
<th>Epidemiology</th>
<th>Most effective</th>
<th>Least effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical format</td>
<td>ARR and RRR</td>
<td>ARR</td>
<td>RRR</td>
</tr>
<tr>
<td>Frequency format</td>
<td>Frequency</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Graphical risk</td>
<td>Ladder</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Pictorial risk</td>
<td>Verbal descriptors, Comparators</td>
<td>Ladder</td>
<td>Frequency</td>
</tr>
<tr>
<td>Qualitative format</td>
<td>Icon arrays, ladder***</td>
<td>Icon arrays, vignettes</td>
<td>Decimals</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Most effective</td>
<td>Frequency</td>
<td>Verbal descriptors</td>
<td>Frequency</td>
</tr>
<tr>
<td>Least effective</td>
<td>Frequency</td>
<td>Verbal descriptors</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Most effective</td>
<td>Icon arrays</td>
<td>Bar chart</td>
<td>Bar chart</td>
</tr>
<tr>
<td>Least effective</td>
<td>Frequency</td>
<td>Ratios</td>
<td>Ratios</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Most effective</td>
<td>Traffic light risk chart</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Least effective</td>
<td>Survival curves</td>
<td>Traffic light risk chart</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>%</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Most effective</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Least effective</td>
<td>Frequency and comparator</td>
<td>Traffic light risk chart</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Epidemiology</td>
<td>No clear conclusion</td>
<td>Traffic light risk chart</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>Most effective</td>
<td>Traffic light risk chart</td>
<td>No clear conclusion</td>
<td></td>
</tr>
<tr>
<td>Least effective</td>
<td>Traffic light risk chart</td>
<td>No clear conclusion</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>ID (year)</td>
<td>Type of risk communication format examined</td>
<td>Most effective</td>
<td>Least effective</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>R45 (2007)</td>
<td>Numerical format: Bar chart, pie chart</td>
<td>Icon arrays</td>
<td>No clear conclusion</td>
</tr>
<tr>
<td></td>
<td>Frequency format: X in 100</td>
<td>Bar chart</td>
<td>Icon arrays</td>
</tr>
<tr>
<td></td>
<td>Graphical risk: Icon arrays</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R46 (2004)</td>
<td>Qualitative format: Verbal descriptors, vignettes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R48 (2006)</td>
<td>% and frequency: Bar chart</td>
<td></td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R49 (2013)</td>
<td>% and frequency: Icon arrays</td>
<td></td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R50 (2003)</td>
<td>% and frequency: X in 10</td>
<td>NNT</td>
<td>Frequency</td>
</tr>
<tr>
<td>R51 (2011)</td>
<td>% and frequency: X in 100</td>
<td>Vignettes</td>
<td>Bar chart</td>
</tr>
<tr>
<td>R52 (2003)</td>
<td>% and frequency: Icon arrays</td>
<td>Verbal descriptors</td>
<td>% combined with verbal descriptors</td>
</tr>
<tr>
<td>R53 (2010)</td>
<td>% and frequency: X in 100</td>
<td>Icon arrays</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R54 (2010)</td>
<td>% and frequency: X in 100</td>
<td>Icon arrays</td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R55 (2008)</td>
<td>% and frequency: Icon arrays</td>
<td></td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R56 (2004)</td>
<td>% and frequency: Bar chart</td>
<td>Icon arrays</td>
<td>Bar charts</td>
</tr>
<tr>
<td>R57 (2009)</td>
<td>% and frequency: 1 in XX and 1:X</td>
<td>Verbal descriptors</td>
<td>Pie Charts</td>
</tr>
<tr>
<td>R58 (2006)</td>
<td>% and frequency: X in 100</td>
<td></td>
<td>Frequency</td>
</tr>
<tr>
<td>R59 (2002)</td>
<td>Vignettes (traffic lights)</td>
<td></td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>R60 (2012)</td>
<td>% and frequency: 1 in XX and 1:X</td>
<td></td>
<td>No clear conclusion</td>
</tr>
<tr>
<td>ID (year)</td>
<td>Type of risk communication format examined</td>
<td>Overall conclusion</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------</td>
<td>--------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Numerical format</td>
<td>Frequency format</td>
<td>Graphical risk</td>
</tr>
<tr>
<td>R60 (2012)</td>
<td>Icon arrays, ladder****</td>
<td>X in 1000</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R61 (2008)</td>
<td>% and frequency</td>
<td>X in 1000</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R62 (2009)</td>
<td>% and frequency</td>
<td>X in 1000</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R63 (2009)</td>
<td>%</td>
<td></td>
<td>Verbal descriptors</td>
</tr>
<tr>
<td>R64 (2007)</td>
<td>Frequency</td>
<td>X in 100</td>
<td>Icon arrays</td>
</tr>
<tr>
<td>R65 (2008)</td>
<td>% and frequency</td>
<td>X in 100 and X in 1000</td>
<td>Icon arrays</td>
</tr>
</tbody>
</table>

*Risk ladder here was referred to as a risk continuum in paper **Risk ladder here refers to a risk thermometer ***Risk ladder here refers to a Paling Perspective Scale ****Risk ladder here refers to the magnifying glass graphic
Effectiveness of risk communication format

To assess the effectiveness of the risk communication format, most studies (n= 53, 82%) asked the study participants to identify the tool they preferred by using a survey-based approach. Qualitative methods of enquiry such as interviews or focus groups were used either alone or in combination with quantitative surveys in 14 studies (22%). The surveys used to assess format effectiveness covered a variety of topics including: self-reported preferences; self-reported behaviour changes; and ability to recall risk information. Almost a third of studies measured respondents’ performance on a numeracy test (n=21, 32%), often using the same three questions (see Figure A3). Details of the different measures of ‘effectiveness’ used by each study can be found in Table A3.

The difference in measures of effectiveness and the difference in risk communication formats compared in each study made it difficult to compare across studies or to analyse the evidence. Table A2 shows which formats the authors concluded were most effective and which were least. Many studies (n=32, 49%) did not achieve consensus on the best risk communication format. Of the numerical methods of presenting information, 11 (17 %) studies found percentages to be the most effective method and 12 (18 %) studies found frequency to be the most effective method.

Ten studies (15%) reported that qualitative descriptions of risk did not aid the understanding of the risk information, and this method appeared to be one of the least successful communication formats. Some studies did conclude that qualitative descriptors were effective when combined with percentages or frequencies (see R4, R8, R22, R52). In terms of graphical approaches to communicate risk, six (9 %) studies reported that bar charts were the most effective communication format (R9, R10, R13, R18, R47, R51). In terms of the pictorial communications of risk, ten studies (15%) reported icon arrays to be the most effective (although R9 and R19 found them inferior). Only two studies reported risk ladders to be the most useful method (R3, R33).
Table A3: Methods for evaluating the effectiveness of risk communication formats used in studies included in the systematic review of risk communication formats

<table>
<thead>
<tr>
<th>ID (year)</th>
<th>Questionnaire</th>
<th>Interview</th>
<th>Focus Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1 (2002)</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R2 (2004)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3 (2012)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R5 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R6 (2010)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R7 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R8 (2007)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R9 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R10 (2006)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R12 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R13 (2001)</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R14 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R15 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R16 (2013)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R17 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R18 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R19 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R20 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R21 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R22 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R23 (2004)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R25 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R27 (2000)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R28 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R29 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R30 (2011)</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>R31 (2012)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>R32 (2009)</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R33 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R34 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R35 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R36 (2003)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R37 (2006)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R38 (2012)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R39 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R40 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R41 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R42 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R43 (2013)</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R44 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID (year)</td>
<td>Questionnaire</td>
<td>Interview</td>
<td>Focus Group</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>R45 (2007)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R46 (2004)</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>R47 (2001)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>R48 (2006)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R49 (2013)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R50 (2003)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R51 (2011)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R52 (2003)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R53 (2010)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R54 (2010)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R55 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R56 (2004)</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>R57 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R58 (2006)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R59 (2002)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R60 (2012)</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>R61 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R62 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R63 (2009)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R64 (2007)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R65 (2008)</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Uncertainty**

The review identified four studies which compared the communication of risk using point estimates with an uncertain range of estimates (an ambiguous risk). Three studies concluded that presenting an ambiguous risk heightened the individual’s concern about the risk in an affect known as ‘ambiguity aversion’ (U1, U2, U4). U4 also found that providing a set of confidence intervals decreased the risk information’s perceived credibility, even in a highly educated sample of respondents. The studies investigating uncertainty matched results in the empirical studies which also found introducing confidence intervals resulted in groups believing the risk was less trustworthy and vague (R47). Conversely, one study (U3) concluded there was no difference between perceived credibility of the risk information presented as a point estimates or a confidence interval, but did report that using either of these methods were superior to the presentation of no risk information at all.

**Non-empirical studies**

Although studies which empirically tested and compared different risk communication formats were of most interest, the search also identified a number of review and overview pieces. The nine reviews which had a systematic search strategy and compared risk communication formats were often very limited in their approach, comparing only a few
formats at a time. For example, one study (S1) looked specifically at frequencies, percentages and epidemiologically terms only, and another (S2) limited their study to graphical communication. The other systematic reviews were limited in their applicability to other health areas, for example one study (S9) looked specifically at the communication of cardiovascular risk.

This structured review also identified twenty-one overview or summary papers. These papers were not deemed sufficient for inclusions in a ‘review of review’ as most were based on opinion, rather than the results of empirical tests. A third of papers (n=7) had only one author (O3, O11, O13, O14, O15, O18, O21) which could indicate the studies had a restricted viewpoint, possibly confined to one individual’s experience.

**Discussion**

This structured review demonstrated a lack of consensus in the risk communication literature on the best risk communication format to use. This finding may relate to the vast amount of possible risk presentation formats and the variety of measures of effectiveness used. A number of studies also concluded that the preferred risk communication format was the one which changed people’s behaviour, but this inference is problematic and care needs to be taken to ensure respondents’ understanding is affected but their preferences are not. Therefore, with such heterogeneity and a lack of consistency in the results it is difficult to draw any substantive conclusions.

The different measures of effectiveness, different comparators and different respondent samples make it challenging to create meaningful comparisons between studies. The discussion here is based on the findings of the empirical studies identified in the review, whilst drawing on some of the opinion pieces. Each risk communication format will be discussed in terms of its advantages and disadvantages, and its potential usefulness in a DCE context.

All different methods are shown in Figure A2 in which the square boxes represent broad categories and the circles represent individual communication formats. The solid lines indicate links and the dotted lines show possible overlaps between categories.
Figure A2: Identified risk communication formats
**Pictorial risk**

The structured review identified a number of different pictorial formats which could aid the communication of risk. One format, a risk ladder involves presenting respondents with the risk of interest next to other higher and lower risks which they may be more familiar with. The ladder can use a log scale or magnifying glasses to incorporate very small and very big risks. Some studies identified in the review used colour coded ‘traffic light’ ladders or thermometers to communicate the risk too. There are standardised risk ladders, such as the Paling Perspective Scale which uses odds that people can relate to (such as the lottery or being struck by lightning) (Moore et al., 2008). Whilst the ladder was found useful in some empirical studies, it was most effective in low-numerate individuals (Keller et al., 2009). There was also debate about whether the comparative risks should be voluntary or imposed, and whether the scales should be linear or logarithmic (Lipkus & Hollands, 1999). Furthermore, the problem of formatting a risk ladder for use in a DCE choice task may prevent successful implementation of the risk communication format.

Icon arrays appeared to be the most well received pictorial communication format which performed best in the empirical studies. Some studies also suggested that icon arrays made it easier for individuals to compare risk magnitudes; fundamental in the trade-off tasks in a DCE. Literature generally supported using blocked icons rather than a scattered approach, as scattering icons appeared to indicate randomness and uncertainty rather than risk (Han et al., 2012).

**Graphs and charts**

Graphical communication of risk is nothing new; Florence Nightingale popularised the use of pie charts in the mid-19th century (Paling, 2003). An advantage of pie charts is that they allow individuals to make a part-whole comparison of the situation. However, the method did not perform well in the review of empirical studies with only one study (R57) identifying it as the best method (and this study used a sample of children). Pie charts also suffered with their ability to communicate small risks where a risk of less than 0.5% can become a very small sector of the image.

Bar charts were the best performing graphical method identified in the review, with many trials reporting successful interpretation of the risk information. One study included in the review found risk was twice as likely to be correctly interpreted when risk was presented as a bar graph, rather than as text containing a percentage (Sprague et al., 2012). Other studies found the horizontal bars took longer to interpret and were better for
communicating uncertainty rather than risk (McCaffery et al., 2012; Schapira et al., 2001). However, the format performed poorly for small numerators such as risks less than 1 in 100 or 1 in 1,000 (McCaffery et al., 2012).

Whilst some empirical studies in the review tested the use of line graphs, these papers generally concluded that this approach was most effective for communicating uncertainty rather than risk. For example, Schapira et al. (2001) used a line graph to show confidence intervals around risks of breast cancer mortality. The review identified a number of studies which had looked at ranges (rather than point estimates) and found them to be inferior in terms of participants’ ability to comprehend and trust the information. Line charts appear to be have limited use in a DCE communicating point estimates.

**Numerical formats**

The structured review identified multiple numerical formats for describing risk information. Percentages are a traditional risk communication format which describe the proportion per hundred. However, for small risks, percentages can be lost with decimal places. Natural frequencies or fractions were another format identified by the structured review. However, frequencies were open to criticisms; such as the numerator and denominator biases. Alternatively, decimal places or odds ratios could be used, but neither of these were any more effective.

**Epidemiological terms**

Some studies used epidemiological terms such as NNT, NNH, RRR and ARR. The epidemiological terms did not perform well in the empirical studies identified by the structured review (and no studies identified this format as ‘best’ when compared to other formats). In addition, it is not clear how well these would perform in a DCE context and whether this would make choices between alternative more confusing.

**Qualitative descriptors**

Qualitative descriptions of risk, where risk is presented as a statement, were met with mixed success. The British National Formulary (BNF) provides guidelines on the appropriate descriptions of probabilities for adverse reactions to drugs shown in Table A4 (BNF, 2015). In cases where the method was found to be effective, it was often combined with numerical information (such as frequencies or percentages).
Other studies have found that qualitative descriptors may indicate uncertainty and cause individuals to assign their own values to each statement (Schmidt, 2004). For DCEs, encouraging simplifying heuristics and respondents to make their own interpretation could induce further ambiguity. Presenting information as illustrated in Table 4.4 could introduce additional information for respondents to remember and could further increase the cognitive burden of the experiment. It may also present additional challenges in the analysis of choice data, for example, choosing an appropriate coding for the quantitative levels of the risk attribute to estimate value trade-offs (such as risk tolerability).

Uncertainty

In DCEs, risk is frequently presented as point estimates with respondents rarely informed about the certainty of the statistic (Harrison et al., 2014). For people who understand risk information, uncertainty may provide them with more useful information. However, the studies in this structured review found individuals were less likely to trust ambiguous risk information. Given the cognitive burden of even a simple DCE, where respondents are required to imagine themselves in a hypothetical scenario considering a health intervention they are likely to be unfamiliar with risk and so the addition of intervals might be an unnecessary complication. Also, the introduction of uncertainty has implications for the coding of levels in the analysis of the DCE data and, possibly, the underlying economic theories which support discrete choice modelling.

Experience and Perceptions

A key finding from the review was that the risk communication format should be tailored to the specific respondent sample. Tailoring is required to take into account the respondents’ experience and prior perceptions, and how this may affect their understanding. This could be because the communication format makes the risk more personable, for example, respondents identifying with the figures in icon arrays and
therefore realising the magnitude presented. However, it could also be important in explaining risks objectively to people who have miscalculated their own susceptibility based on family experience. In the context of DCEs this is particularly important as respondents should base their choices on the risk information presented rather than their own inferred probabilities.

**Numeracy and mathematical competency**

The studies identified in the review used a range of samples and measured understanding of risk in various ways: some studies used interviews; a smaller number of studies used ability to recall information; and others tested the formats using mathematical questions. Of the studies using mathematical questions to assess the numeracy skills of their sample (regardless of the outcome measures), there were some questions commonly occurring. The original questions by Schwartz et al. (1997) are shown in Figure A3. Although there were some slight variations across other studies in the actual terminology used to frame these numeracy questions, they were qualitatively similar to those suggested by Schwartz et al. (1997).

**Figure A3: Standardised numeracy questions**

1) Imagine that we rolled a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even?

2) In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize is 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket to BIG BUCKS?

3) In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car?

*Source: Schwartz et al. (1997)*

**Evaluating risk communication formats**

The challenge of evaluating the most effective risk communication format is the balance between achieving what is attractive and appealing, what changes behaviour through manipulation and what actually improves an individual’s ability to reason with risk. As Table A3 shows, multiple methods were used to measures format effectiveness. In the case of DCEs, the risk communication format should improve the individual’s ability to make an informed decision but this outcome is difficult to measure quantitatively. In the original proposal of work, this PhD intended to use qualitative interviews to appraise the framing
methods identified in the review. However, the usefulness of qualitative interviews alone was viewed to be minimal given the multiple outcome measures and mismatching between self-reported appeal and objective numerical performance. Therefore proceeding with a DCE employing risk communication formats identified by the structured review was thought to be the most appropriate approach.

**Strengths and limitations**

This review identified a number of ways of communicating risk information. The search strategy was broad in its range of databases but used specific terms, and this could be seen as a limitation. However, the specific search strategy and strict exclusion criteria facilitated a rapid review of a large research area. A lack of second reviewer and extractor may have resulted in missed papers or risk communication formats. However, the number of new framing formats decreased considerably towards the end of the extraction process suggesting data saturation was achieved. There were also many studies which could not conclude which method performed best. No studies included in this review compared all the identified formats which prevent a solid conclusion on the ‘best’ or the ‘worst’ risk communication formats.

**Conclusion**

The results of this review uncovered that there is a fine line between providing sufficient information for respondents to understand the risk and overwhelming them with too much. Training materials for respondents should be presented at a level that is accessible to all but should not be so elementary that respondents will ignore it. A prescriptive approach to the DCE, with extensive piloting and revisions, should therefore be conducted to ensure materials are appropriate for respondents.

The identified empirical studies also highlighted the importance of assessing numeracy skills on a respondent’s ability to interpret probabilistic information. Preformed ideas about perceived risk based on experience were also identified as a challenge to effective communication of risk information. In understanding how respondents interpret risk in a DCE, numeracy, experience and perceptions could be useful covariates to collect in supplementary questions in a DCE study.

The review identified icon arrays as the risk communication format with the most empirical support which would make it an appropriate comparator to current practice in DCE literature which are percentages only (Harrison et al., 2014). However, no empirical
study investigated the use of icon arrays in a trade-off task so it is unclear whether they would be superior in a DCE context.

**Bibliography**


Appendix 3.2: List of studies included in the structured review of risk communication formats

<table>
<thead>
<tr>
<th>ID</th>
<th>Full reference</th>
</tr>
</thead>
</table>


Timmermans DRM, Ockhuysen-Vermey CF, and Henneman L. 2008 Presenting health risk information in different formats, the effect on participants’ cognitive and emotional evaluation and decisions. Patient Education and Counseling, 73, 443–7.


O21  Zikmund-Fisher, B.J., (2013). The right tool is what they need, not what we have, a taxonomy of appropriate levels of precision in patient risk communication. Medical care research and review, MCRR, 70(1 Suppl), p.37S–49S.


Risk as an Attribute in Discrete Choice Experiments: A Systematic Review of the Literature

Mark Harrison, Dan Rigby, Caroline Vass, Terry Flynn, Jordan Louviere & Katherine Payne

The Patient - Patient-Centered Outcomes Research
ISSN 1178-1653
Volume 7
Number 2
Patient (2014) 7:151-170
DOI 10.1007/s40271-014-0048-1
Risk as an Attribute in Discrete Choice Experiments: A Systematic Review of the Literature

Mark Harrison · Dan Bigby · Caroline Vass · Terry Flynn · Jordan Louviere · Katherine Payne

Published online: 25 February 2014
© Springer International Publishing Switzerland 2014

Abstract
Background Discrete choice experiments (DCEs) are used to elicit preferences of current and future patients and healthcare professionals about how they value different aspects of healthcare. Risk is an integral part of most healthcare decisions. Despite the use of risk attributes in DCEs consistently being highlighted as an area for further research, current methods of incorporating risk attributes in DCEs have not been reviewed explicitly.

Objectives This study aimed to systematically identify published healthcare DCEs that incorporated a risk attribute, summarise and appraise methods used to present and analyse risk attributes, and recommend best practice regarding including, analysing and transparently reporting the methodology supporting risk attributes in future DCEs.

Data Sources The Web of Science, MEDLINE, EMBASE, PsycINFO and Econlit databases were searched on 18 April 2013 for DCEs that included a risk attribute published since 1995, and on 23 April 2013 to identify studies assessing risk communication in the general (non-DCE) health literature.

Study Eligibility Criteria Healthcare-related DCEs with a risk attribute mentioned or suggested in the title/abstract were obtained and retained in the final review if a risk attribute meeting our definition was included.

Study Appraisal and Synthesis Methods Extracted data were tabulated and critically appraised to summarise the quality of reporting, and the format, presentation and interpretation of the risk attribute were summarised.

Results This review identified 117 healthcare DCEs that incorporated at least one risk attribute. Whilst there was some evidence of good practice incorporated into the presentation of risk attributes, little evidence was found that developing methods and recommendations from other disciplines about effective methods and validation of risk communication were systematically applied to DCEs. In general, the reviewed DCE studies did not thoroughly report the methodology supporting the explanation of risk in training materials, the impact of framing risk, or exploring the validity of risk communication.

Limitations The primary limitation of this review was that the methods underlying presentation, format and analysis of risk attributes could only be appraised to the extent that they were reported.

Conclusions Improvements in reporting and transparency of risk presentation from conception to the analysis of DCEs are needed. To define best practice, further research is
needed to test how the process of communicating risk affects the way in which people value risk attributes in DCEs.

**Key Points for Decision Makers**

- Discrete choice experiments (DCEs) in health often use risk attributes to value risk-benefit trade-offs and potentially the perceived value of risk reduction.
- People struggle to understand concepts of risk and probability, and process and distinguish between differences in risk.
- Improvements in reporting and transparency of risk presentation from conception to the analysis of DCEs are needed.
- Recommendations about risk presentation from other fields need to be adopted by researchers developing DCEs.
- Experimental studies to understand the impact of different methods of risk presentation/communication are urgently needed.

1 Background

Risk is a ubiquitous component of healthcare and decisions made by patients and healthcare professionals about commencing or continuing treatments that are shaped by perceived benefits and risks. Discrete choice experiments (DCEs) allow trade-offs between health and non-health outcomes and processes to be valued. Including a cost attribute allows indirect elicitation of willingness-to-pay (WTP) values [1]. In theory, including a risk attribute allows valuation of the balance between risk and benefits (risk-benefit trade-offs). In the decision theory literature, risks (prospects with known probabilities associated with possible outcomes) are distinguished from uncertainties (prospects with unknown probabilities of possible outcomes) [2]. However, significant potential difficulties in presenting risk information are recognised and if probability information is not well understood, not well presented or simply ignored then this limits the validity, usefulness and applicability of the results. A substantial, growing literature has explored the way humans perceive and interpret risk and probability information [3], and an extensive literature about improving risk communication is emerging [4–6].

Presentation of risk information has been studied in the contingent valuation (CVM) literature. Findings that WTP for risk reductions from CVM were inconsistent with economic theory prompted methodological research into risk communication, close attention to risk communication methods, and rigorous checks of validity [3, 7]. The recommendations were that CVM could provide theoretically valid estimates of WTP for risk reduction if certain methods were applied and tested for validity against standard economic theory; WTP should be positively associated with a magnitude of risk reduction (or change), proportional to the size of risk reduction, and when the baseline risk is small the effect of the baseline risk (and changes in this risk) on WTP should be small [3, 7]. Methods of enhancing risk communication is CVM studies to achieve results consistent with economic theory centred on using visual aids (icon/dot arrays, risk ladders using a logarithmic scale, analogies) to set the risk in context [3, 7].

Detailed reviews of the application and methodological basis of DCEs in healthcare consistently highlight including risk as an attribute as an important area for further research, perhaps indicating a lack of progress in this area [1, 8–10]. To date, no review has explicitly summarised the methods used to incorporate risk as an attribute in DCEs. The aim of this paper was to supplement previous reviews and recommendations [11, 12] by identifying published DCE studies in healthcare that incorporated risk as an attribute and appraising the methods underlying presentation, format and analysis of risk attributes. The objective is not to criticise the authors of current healthcare DCEs, nor the methods used to date, but to summarise the current methodology, highlight potential areas that might be informed by developments in other fields, and make recommendations to improve risk communication in healthcare DCEs. The key objective is to provide a narrative review summarising the current state of incorporating risk in healthcare DCEs and to move towards recommendations on best practice for future DCEs or aspects of risk presentation in DCEs requiring further methodological exploration. This key objective was met by also completing a rapid review of the risk communication literature.

2 Methods

This study used two reviews: (1) a systematic review to identify relevant published work related to valuing a healthcare intervention using a DCE with risk included as an attribute; and (2) a rapid review to identify examples of good practice in risk communication in the general (non-DCE) health literature.

2.1 Systematic Review of Risk in Discrete Choice Experiments (DCEs)

Risk attributes in DCEs in healthcare potentially aim to reflect the uncertain consequences of treatments. Typically,
Risk may be defined as being knowable (i.e. the different potential outcomes have a quantifiable probability of occurring), whereas uncertainty cannot be quantified [13]. This review takes an inclusive approach to the definition of risk, consistent with the decision theory literature, where risk attributes are defined as those presenting quantitative estimates of the probability associated with an uncertain outcome (either unwanted consequences or forgone benefits). However, we also include DCEs using qualitative descriptions of the likelihood of the unwanted prospect (or harms) to avoid potentially excluding relevant studies.

The electronic search strategy (Table 1) was derived from several sources, including two recent reviews in the area [1, 10], and a trained librarian (Ingram M, personal communication). Studies including a risk attribute were identified at the manual screening stage. The following databases were searched (on 18 April 2013): Web of Science, MEDLINE, EMBASE, PsyCINFO and Econlit. The search strategy was validated by checking that references from the two recent reviews and other studies known to the team were captured. Previous systematic reviews of DCEs in healthcare identified two DCEs published before 1995, neither of which included a risk attribute [1]. Therefore, the time horizon was restricted to studies published since 1995.

All abstracts were screened independently to identify DCEs with risk attributes by three reviewers (MH and KP, as well as Stuart Wright), who met to discuss any disagreement. The inclusion criteria were healthcare related, DCEs (respondents make a clear choice between two or more alternative technologies/services comprising several attributes), risk—or an attribute consistent with our inclusive definition of a risk attribute—is mentioned in the title/abstract. Health risks associated with non-healthcare products (e.g. food) were excluded. Conference abstracts and other examples of the non-peer reviewed literature, non-English articles and opinion/review/protocols/conference abstracts were excluded. Following the initial screening of abstracts, full copies of papers were obtained and retained in the final review if a risk attribute meeting our definition was included. Papers that framed the risk in certain terms (e.g. present/absent) were excluded.

Extracted data were tabulated. A published framework for critically appraising DCEs was used to summarise the quality of reporting in the DCEs [8]. In addition, the format, presentation and interpretation of the risk attribute were summarised using guidance from the CVM literature [7]. The review focussed on collating the following: background information/explanation of risk presented to respondents; framing of risk (description of risk attributes and format used to communicate changes in risk); stated assumptions about risk; analysis and interpretation of risk attributes (checks of linearity of risk preferences, sensitivity to scope, or risk attitude; calculating marginal rates of substitution [MRS] or WTP for risk reduction).

2.2 Rapid Review of Risk Communication Tools

Given the substantial and diverse literature that describes and/or evaluates different tools for the effective communication of risk in healthcare, a focussed rapid review was conducted to identify approaches to healthcare risk communication. The review was focussed by identifying studies that (1) tested any risk communication tool that aimed to aid people’s understanding of quantitative information; (2) were related to a healthcare risk; and (3) were published in peer-reviewed journals from January 2000 onwards and written in English. Studies that looked at public health risks, such as nationwide vaccines, and studies that did not use written communication, such as videos or games, were excluded. The search strategy was devised with assistance from a trained librarian (Ingram M, personal communication) and was validated by checking that references from key studies known to the team were captured. The search strategy included the terms Title = (risk*) AND Title = (communicat* or format*) AND Topic = (aid* or present* or display*). The following databases were searched (on 23 April 2013): Web of Science, MEDLINE, EMBASE, PsyCINFO and Econlit.

An initial title and abstract screen to select relevant studies was conducted. If studies could not be rejected with certainty, the whole article was retrieved and reviewed. After retrieving all the full papers, the studies were initially categorised into four categories: empirical studies testing risk communication tools; systematic review of risk communication tools; papers that provided an overview; and studies that looked specifically at communicating uncertainty. Data extraction focused on describing the risk communication tool and whether a definitive conclusion was drawn regarding the relative effectiveness of each tool. These data were then summarised in a narrative review.
3 Results

3.1 Systematic Review of Risk in DCEs

Overall, 117 papers were identified and included. Figure 1 summarises the study identification and inclusion process. Few studies published before 2002 included a risk attribute, but thereafter the number increased to over 20 per year in 2011 and 2012 (Fig. 2). As a proportion of the total number of DCEs conducted [10], the growth in those with a risk component has been less dramatic. The included papers covered a range of healthcare settings, most commonly cancer (n = 27, 23%). Most DCEs investigated preferences for pharmacological treatments (n = 70, 60%), followed by screening and diagnostics (n = 15, 13%). Fifty-nine (50%) studies reviewed included a monetary attribute and, of these, most (n = 43, 73%) calculated WTP (usually n = 40 [68%] for risk reduction). Two (2%) studies calculated the value of a statistical life [14, 15]. The mean number of attributes in the DCEs was six (range 3–12) and a mean of two (range 1–7) risk attributes were included. In three (3%) DCEs, all attributes were risk attributes [16–18].

The results of the review are categorised into aspects of DCEs, presented in the order in which the reader would encounter each aspect when designing, framing, administering, analysing and interpreting DCEs. Each aspect is presented and reviewed in turn.

3.1.1 Explanatory and Background Training Materials

The reporting of background training materials presented was generally insufficient to assess the nature of explanatory material offered to respondents. Seventy-one (61%) of the 117 identified studies reported providing background training information to participants, but only 27 (23%) described this information in any detail. When described, background material included detail about the intervention or procedure for completing the DCE, although it was unclear whether any background on risk was provided. Three (3%) studies provided full access to the entire survey materials in an online supplementary file [19–21], whilst others summarised information provided to respondents in a table [22] or written descriptions of multiple sections which helped respondents understand their risk attitude. Further, two (2%) studies reported providing a risk tutorial to enable respondents to interpret risk [23, 24].

3.1.2 Choice Question Format

Respondents were given an opt-out option with the choice sets in 41 (35%) studies. Seventeen (15%) studies offered a status quo opt-out, and 11 (9%) defined the risk associated with this option to the respondent in the choice set, either as population risk values [19, 25–33] or the respondent’s current choice and risk calculated using information collected during earlier sections of the questionnaire [23]. Five (4%) studies allowed the respondents to consider their own current condition [14, 34–37], for example their child’s current risk [36].

3.1.3 Communication of Risk Attributes

3.1.3.1 Presentation Risk attributes were mostly presented quantitatively (n = 72, 62%), using frequencies (n = 23, 32%), percentages (n = 27, 38%) or a combination of both (n = 22, 31%). In 13 (11%) studies, risks were presented with varying denominators, although none presented risks of equal magnitude using different denominators (ratio bias). Six (5%) studies varied denominator for a risk within the same attribute [26, 35, 38–41]; for example comparing a 1 in 5 with a 1 in 10 chance [40]. However, these estimates were supplemented with either qualitative descriptions (e.g. “minimal risk [1 in 150,000]” [39]), percentages [40], or both (e.g. “moderate (1 in 5 or 20%)” [33]). Five (4%) studies used varying denominators between attributes [42–46], including one that presented two different side effects using numbers of
In all, 279 episodes but with a varying timeframe for occurrence (one in 2 years, one in 3 years) [44] and a further seven (6 %) used different denominators within and between attributes [26, 40, 47-51]. Nineteen (16 %) studies used a purely qualitative approach, suggesting that these attributes should be viewed as 'uncertain' instead of 'risks'; three (3 %) studies in haemophilia took an approach of describing risk in the context of existing (and assumed understood) products, e.g., risk of viral infection from human plasma-derived or recombinant products [52, 53], or current level of risk [54]. Twenty-six (22 %) studies combined qualitative and quantitative methods to present risk. In some DCEs, qualitative and quantitative methods were used to describe different risk attributes, and in others within the same attribute, seemingly to aid communication (e.g. 'low, 1 in 20 (0 %)') [47]. Descriptions used for risks varied between studies; for example, risks considerably lower than 1 % were described as 'minimal' (1 in 150,000) [39] and 'very small' (1 in 10,000) [55], a risk of around 1 % was described as 'very low' and 10 % was described as 'low' [38], whereas others use 'low' for a risk <10 % [56].

Elsewhere, risks in the region of 20 % were described as 'moderate' [38] and 'medium' [56], and risks beyond 30 % were described as 'high' [56]. Where studies used purely qualitative terms such as 'small', the risk magnitudes supporting these descriptions were unclear [57, 58]. Absolute risks (n = 88, 73 %) were presented more often than relative risk (n = 8, 7 %), although five (4 %) studies included both [14, 27, 59-61]. Only two of eight studies using relative risk alone gave a baseline/underlying risk estimate to place the difference/change in overall risk in context [15, 33]. Risk attributes were virtually always presented as point estimates without variability or a range around each defined attribute level. One exception presented a range of values (e.g. $<10$, $10-30$, $>30$ % [56]), although it is unclear whether this represents variability or an interval description. Others used risk levels ('$<1$ %' and 'about 5 %') [62], and thresholds ($\geq2$, $\geq5$ or $\geq10$ in 100 patients) [63]. Other qualitative studies qualified the point estimates using words like 'almost' or 'about' to convey a lack of certainty [64, 65]. In one study, uncertainty related to two quantitative risk attributes (bone or kidney damage) was introduced using a qualitative attribute describing whether they could be successfully treated [66].

Visual aids were used to support communication of risk attribute information in 27 (23 %) studies. The most common visual risk presentation methods were risk icon/dot arrays (n = 25, 21 %). Other methods included bar charts (n = 3, 3 %), pictograms, and risk ladders that represent the risk in the DCE alongside other, more familiar, risks (Table 2). Photographs were also used to help describe risk attributes [29, 40, 55].

### 3.1.3.2 Framing

Risk was most often framed negatively as a consequence of treatment or an intervention (n = 102, 87 %) rather than positively as something to be avoided (n = 26, 22 %). Eleven (9 %) studies included a mix of positively and negatively framed risk attributes. Negative framing of risk most often concerned drug-related events or reactions, death, or adverse events/complications following treatment. Positively framed attributes concerned risk reductions and adverse events/complications avoided. Howard and Salkeld [45] explored the impact of positive or negative framing of three attributes related to the detection...
<table>
<thead>
<tr>
<th>Year</th>
<th>Author (ref.)</th>
<th>Sample</th>
<th>Risk</th>
<th>Absolute/relative risk</th>
<th>Graphical risk</th>
<th>Linearity in risk tested</th>
<th>Sensitivity to scope scale</th>
<th>WTP for risk reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>Rachele et al. [50]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Lloyd et al. [40]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Aschcroft et al. [74]</td>
<td>HP</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Sexton et al. [122]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Johnson et al. [78]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>Lloyd et al. [47]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Hauber et al. [63]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Johnson et al. [79]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Eisen et al. [42]</td>
<td>PUB</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Hauber et al. [143]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Svinburn et al. [46]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Hauber et al. [144]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>van Houtten et al. [138]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Mohamed et al. [145]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Pereira et al. [51]</td>
<td>DM</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Hauber et al. [146]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Gao et al. [25]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Kant et al. [24]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Hodgkins et al. [147]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Asden et al. [148]</td>
<td>HP</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Lathia et al. [149]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Hauber et al. [150]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Shackley et al. [71]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Hall et al. [48]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Sallhed et al. [43]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Assisdo et al. [151]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>McTiggen-Cowan et al. [44]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>Tinetti et al. [16]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>Tinetti et al. [18]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Guimaraes et al. [152]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Howard et al. [45]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>de Bakker-Grob et al. [49]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>Photogram, bar chart</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2011</td>
<td>Oyen et al. [19]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Sweeting et al. [20]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Sclame et al. [26]</td>
<td>PT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Author (ref.)</td>
<td>Sample</td>
<td>Risk</td>
<td>Absolute/relative risk</td>
<td>Graphical risk</td>
<td>Linearity in risk tested</td>
<td>Sensitivity to scope scale</td>
<td>WTP for risk reduction</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>--------</td>
<td>------</td>
<td>------------------------</td>
<td>----------------</td>
<td>------------------------</td>
<td>---------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>2011</td>
<td>Howard et al. [151]</td>
<td>FT</td>
<td>Freq.</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Faggioni et al. [154]</td>
<td>FT/HP</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Laba et al. [22]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Bridges et al. [72]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Gudjónsson et al. [155]</td>
<td>Parent</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Pigorini et al. [21]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Pigorini et al. [17]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>de Bekker-Grob et al. [27]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR and RR&lt;sup&gt;str&lt;/sup&gt;</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Augustin et al. [156]</td>
<td>FT</td>
<td>Freq.</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Vennix et al. [28]</td>
<td>PT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>Bryan et al. [157]</td>
<td>PT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Bryan et al. [158]</td>
<td>PT</td>
<td>%</td>
<td>AR</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Telser and Zweifel [81]</td>
<td>PUB</td>
<td>%</td>
<td>RR</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Weston and Fitzgerald [62]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Bishop et al. [139]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Lewis et al. [160]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>Lewis et al. [161]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Telser and Zweifel [82]</td>
<td>PUB</td>
<td>%</td>
<td>RR</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Walzer [36]</td>
<td>PT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>de Bekker-Grob et al. [162]</td>
<td>FT</td>
<td>%</td>
<td>RR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Bunge et al. [163]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Watson et al. [64]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>de Bekker-Grob et al. [165]</td>
<td>PT</td>
<td>%</td>
<td>RR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Knujijhar et al. [166]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Hol et al. [167]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Turbica and Pastore [59]</td>
<td>HP</td>
<td>%</td>
<td>AR and RR&lt;sup&gt;str&lt;/sup&gt;</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Nayaradou et al. [60]</td>
<td>PUB</td>
<td>%</td>
<td>AR and RR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Dumas et al. [92]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Wirtzko et al. [158]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Flood et al. [169]</td>
<td>Parent</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Schneidem et al. [65]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Regier et al. [50]</td>
<td>Parent</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Moehl [31]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Dammann et al. [130]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Tong et al. [171]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Song et al. [172]</td>
<td>FT</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Boeri et al. [23]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Author (ref.)</td>
<td>Sample</td>
<td>Risk</td>
<td>Absolute/relative risk</td>
<td>Graphical risk</td>
<td>Linearity in risk tested</td>
<td>Sensitivity in scope/scale</td>
<td>WTP for risk reduction</td>
</tr>
<tr>
<td>------</td>
<td>---------------</td>
<td>--------</td>
<td>------</td>
<td>-------------------------</td>
<td>---------------</td>
<td>------------------------</td>
<td>--------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Qualitative risk attributes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>Mantovani et al. [54]</td>
<td>PT/AIP</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>Lee et al. [52]</td>
<td>HP</td>
<td>Description and Y/N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>Scalone et al. [53]</td>
<td>PT/AIP</td>
<td>Y/N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>Müller et al. [70]</td>
<td>PT</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Marang-van de Meene et al. [173]</td>
<td>PT</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Watson et al. [174]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Bijlenga et al. [175]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Bijlenga et al. [176]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Halmi et al. [177]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Müller and Husking [178]</td>
<td>PT/AIP</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Trulik et al. [32]</td>
<td>PT</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Mansmajd et al. [179]</td>
<td>PT</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Kinsler et al. [180]</td>
<td>PT</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>King et al. [181]</td>
<td>PT</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Burnett et al. [182]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Bridges et al. [183]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Lee et al. [184]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Wintink et al. [185]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Zimmermann et al. [186]</td>
<td>PUB</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative and qualitative risk attributes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Lloyd et al. [38]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2009</td>
<td>Hamer et al. [66]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Array</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2010</td>
<td>Johnson et al. [80]</td>
<td>PT</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Array</td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2010</td>
<td>Cameron et al. [18]</td>
<td>PUB</td>
<td>Freq. and %</td>
<td>AR and RR$^{RL}$</td>
<td></td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2011</td>
<td>Mohamed et al. [187]</td>
<td>PUB</td>
<td>Freq. and %</td>
<td>AR</td>
<td>Array</td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2005</td>
<td>Tsuge et al. [15]</td>
<td>PUB</td>
<td>Freq.</td>
<td>RR$^{RL}$</td>
<td>Array</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2006</td>
<td>Epelde et al. [55]</td>
<td>PT/AIP</td>
<td>Freq.</td>
<td>AR</td>
<td>Photograph</td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2007</td>
<td>Osa et al. [99]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td></td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2009</td>
<td>Petrou and McIntosh [388]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2011</td>
<td>Shaheen et al. [61]</td>
<td>PT/AIP</td>
<td>Freq.</td>
<td>AR</td>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2011</td>
<td>Fiebig et al. [35]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2013</td>
<td>Sadique et al. [189]</td>
<td>PUB</td>
<td>Freq.</td>
<td>AR</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2002</td>
<td>Artsides et al. [190]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2004</td>
<td>Watson et al. [191]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2005</td>
<td>Lee et al. [96]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td></td>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2005</td>
<td>Frumel et al. [192]</td>
<td>PUB</td>
<td>%</td>
<td>AR</td>
<td>Bar chart</td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
and treatment of cancer. People seemed more willing to risk potential harms and had a higher WTP to detect cancer than to reduce the chance of missing a cancer, suggesting evidence of framing effects. This is consistent with previous findings that people were more willing to pay for additional lives ‘saved’ rather than a reduction in lives ‘lost’ [67], and variance heterogeneity whereby people have different levels of certainty for losses and gains (see below) [68, 69].

The timeframe over which the risky ‘event’ should be considered by the respondent was stated explicitly in 63 (54 %) studies. Fifteen (13 %) study provided information on risk latency, the delay between the decision and experiencing the ‘risk’, ranging from immediate to 10 years [15, 57, 70]. In 11 (9 %) studies the risk appeared to be immediate due to the nature of the intervention and risk (e.g., during/following surgery) [56, 71, 72]. The remaining studies provided risk timeframes but did not specify any latency.

### 3.1.4 The Study Sample

Study samples most often comprised patients (n = 66, 56 %), parents or carers of patients (n = 4, 3 %), or healthcare professionals (n = 6, 5 %), suggesting most preferences were ex-post. Ex-ante preferences also were elicited from the public (n = 28, 24 %). Twelve (10 %) studies used multiple respondent types to compare patient preferences with other groups to better understand shared decision making or ex-post with ex-ante preferences.

### 3.1.5 Inclusion of Supplementary Questions

Most DCEs included supplementary sections seeking to identify respondent characteristics or knowledge of the intervention. None reported including questions to assess numeracy or literacy. Based on the reporting of the studies, respondent risk behaviour was not generally explored in the DCEs. Only Tsuge et al. [15] reported exploring risk attitudes, focusing on respondents’ perceptions of risks (questions to gauge responsibility for the risks used in the DCE, and risk avoidance behaviour). Goto et al. [73] assessed the knowledge that respondents had about associations between smoking and a range of diseases. One study reported free text responses on the decision-making process made by clinicians whilst completing the DCE; clinicians felt that they focused on longer-term risks whilst patients focussed on short-term benefits [74].

### 3.1.6 Analysis of Choice Data

Most studies (n = 53, 45 %) assumed a linear additive indirect utility function, implying that only additive main effects matter and were estimated. However, 25 (21 %) studies stated that interaction terms were included to allow multiplicative effects.
Interactions can be used to estimate two-way and higher order interactions between attributes. Ten (9%) studies interacted two risk attributes, e.g. the size of risk reduction with the type [15]. Two (2%) DCEs interacted two risk attributes (one mild, one serious AE) in a design allowing all two-way interactions between attributes to be estimated [48, 49]. Neither found a significant interaction but highlight the need to ensure face validity of risk attribute interactions so that significant interactions can be explained meaningfully.

Sensitivity to scope (or scale) of risk/risk reductions or testing for concave increases in WTP (diminishing marginal returns to scale) with increasing benefits (e.g. risk reduction) [75, 76] was reported to have been evaluated in 11 (9%) studies. The terms ‘scale’ and ‘scope’ tend to be used interchangeably [77] but some studies draw distinctions between sensitivity to the size of the benefit (scale) (e.g. WTP for a 50% compared with a 25% risk reduction) with scope (e.g. comparing WTP for a 25% risk reduction of heart disease with a 25% risk reduction from heart disease and diabetes [75]). Several studies tested the rate at which preferences/WTP increase (i.e. proportional or concave), although a stronger test would be whether preferences are strictly proportional to the risk level. Thus, tests of sensitivity to scope overlap with tests of linearity in risk. We found evidence in 36 (31%) studies of linearity in risk being explored, most commonly using dummy coding to account for risk/risk reductions of unequal size with interpretation of the proportionality of coefficients/MRS [50, 55] and/or graphical presentation of moving between the levels of the risk attribute. Several studies used piecewise-linear extrapolation to estimate preferences between discrete risk attribute levels [63, 78–80]. Other approaches included coding attributes for differences between levels, using quadratic terms to identify non-linear relationships, polynomial approximations to test the proportionality of WTP for risk reductions with the magnitudes of risk reduction [15], and statistical tests of linearity of coefficients [81, 82].

3.2 Review of Risk Communication Tools

The search of all databases revealed 1,207 possible hits. After the removal of non-peer-reviewed material and duplicates, 300 titles and abstracts were reviewed. In total, 215 full articles were retrieved and 99 papers were included in the final review (electronic supplementary material [ESM] Appendix 1).

Of the 99 included studies, 65 (66%) were empirical tests of risk communication tools, 21 (21%) were overviews of the risk literature but did not report a search strategy, nine (9%) were systematic reviews with a structured search strategy, and four (4%) looked at presenting uncertainty (for example, the use of confidence intervals as opposed to point estimates). ESM Appendix 2 summaries the 65 studies that reported an empirical test of a risk communication tool. These 65 studies most commonly explored numerical presentation of risk information such as percentages, fractions, decimals or ratios (n = 55, 85%). Qualitative descriptions and verbal analogies were investigated in 34 studies (52%). Other methods included pictorial presentations, such as icon arrays (n = 29, 45%), risk ladders (n = 10, 15%) and graphical information (n = 15, 23%). The criteria to quantify the ‘effectiveness’ of the tool as a risk communication strategy was measured using a numeracy test (n = 21, 32%) or by asking the study participants which tool they preferred using qualitative methods of enquiry or a non-numerical questionnaire (n = 47, 72%).

A number of studies stated clear conclusions that risk ladders, icon arrays and bar charts were the best way to communicate risk. Of the numerical methods of presenting information, 11 (17%) studies found percentages to be the most effective method [6, 83–92] and 12 (18%) studies found frequency to be the most effective method [93–104]. For graphical representations of risk, six (9%) studies found bar charts to be the most effective communication method [105–110]. Ten studies reported that qualitative descriptions of risk did not aid people’s understanding of the information, and this method was one of the least successful communication tools [83, 85, 98, 100, 111–116].

4 Discussion

The use of risk attributes in DCEs is common. Evidence from the identified reviews suggests that researchers conducting DCEs with risk attributes are often using the appropriate basic risk information (e.g. absolute risks with denominators), and visual aids such as icon/arrays. However, there was little evidence that indicated that there was a consistent approach to communicating/framing risk information based on evidence from the risk communication literature in DCEs, to the extent as in related fields such as CVM [3, 7]. However, this may, in part, be due to a lack of reporting of methods used to support risk presentation because of journal word count restrictions.

Although there was a variety of (occasionally conflicting) conclusions to the studies included in the rapid review, some substantive conclusions could be drawn. The rapid review of risk communication methods found evidence supporting more sophisticated methods of presenting risk information through the use of graphical or pictorial images, the use of icon arrays and risk ladders. The use of icon arrays in DCEs with risk attributes is increasing over time.

The implications of the findings are now discussed under the headings under which elements of the inclusion
Table 3  Key recommendations

<table>
<thead>
<tr>
<th>Aspect of DCE</th>
<th>Key recommendation(s)</th>
</tr>
</thead>
</table>
| Explanatory and background training materials | 1. Should be documented and available online (and maintained) in journals in which they appear. The quality of explanatory material should be subject to testing to the same extent as the statistical properties of the design and analysis.  
2. Should define and place risk in context. The amount of explanatory/background information should be equivalent for all attributes to avoid overstating the importance of the risk attribute.  
3. The numerical and risk literacy of respondents should be assessed during the piloting of the DCE. Strategies to overcome problems with numerical literacy should be used.  
4. Appropriate visual aids, diagrams and other methods, such as information acceleration should be considered. The impact of the use of alternative means of risk communication should be tested formally. |
| The choice question format | 5. Opt-out options may be required in DCEs to reflect real life situations. Researchers need to carefully define the research question and consider different ways of framing risk and uncertainty information in DCEs.  
6. The meaning of status quo and no treatment opt-outs in terms of risk should be clear and conveyed to respondents where possible. If these are not known ex-ante, there is a need to elicit how respondents defined the underlying risk and incorporate that into the analysis.  
7. Differences in error variance between researcher- and respondent-defined opt-out attribute levels should be tested. |
| Framing and description of attributes | 8. There is a need for greater interdisciplinarity collaboration with experts from other fields such as psychology when framing and describing attributes to benefit from their expertise.  
9. There is a need for empirical evidence to indicate whether good practice in presenting/communicating risk from the risk communication and CVM literatures should be adopted. There is a need for systematic testing that these methods of presenting risk information, which may change responses, improve the communication of risk.  
10. The method of presenting risk should be validated using tests such as internal magnitude tests, tests of sensitivity to probability, and responses to changes in risk [3, 7].  
11. A consistent format, including a consistent denominator, should be used when communicating risks and benefits [4, 5]. |
| The study sample | 12. Where possible, information on the respondents' knowledge of, and experiences with, the risk presented should be sought and included as potential determinants of preference heterogeneity in the analysis to aid interpretation of preferences. |
| Data collection | 13. Online data collection, which allows more flexible methods of risk presentation to be incorporated and developed, such as visual, audio and other technological aids, should be considered. |
| Inclusion of supplementary questions | 14. Supplementary questions should be used to better understand respondent characteristics such as attitude to risk, numeracy, and risk literacy that can affect trading behaviour and preferences for risk-benefit trade-offs. |
| Analysis of choice data | 15. Careful attention should be paid to the assumed functional form for the way that risk attributes enter into utility functions, and the design and analysis should be carefully planned to support the assumptions made.  
16. Assumptions made about attitudes to risk and how risk has been interpreted should be tested formally at the analysis stage. It is important to consider whether there is evidence of preference heterogeneity, which may have been driven by attitudes to risk, numeracy and risk literacy. |

DCE discrete choice experiment

of risk within DCEs were characterised. Table 3 lists key recommendations relating to each section. Recommendations 3, 4 and 9 highlight areas for future research required before best practice can be defined; the remainder relate to the need for transparent reporting of study design and approaches to include risk.

4.1 Explanatory and Background Training Materials

Researchers designing DCEs need to ensure that respondents receive consistent, correctly communicated information which is then assumed to be used in making choices [117]. For respondents to make informed, reliable and incentive-compatible choices and trade-offs, particularly when trading off risk against non-risk attributes, they must be sufficiently interested to pay attention to details and understand the trade-offs presented [118, 119]. If risk is not effectively communicated to respondents, the result may be large variances around responses that preclude meaningful elicitation of preference information. The use of step-by-step descriptions or numerical data relating to risk may be useful as a training tool [4]; however, there
was little evidence that this technique has been used. The potential for utilising this kind of approach is increasing as online or computerised administration allows more flexible presentation of information.

The evaluation of the quality of background or explanatory materials was restricted by the lack of availability of materials to readers of published DCEs, although encouragingly some recent studies published their entire surveys in online appendices [10–21].

4.2 Choice Question Format

The decision to include ‘opt-outs’ or ‘no choice’ options should be based upon the research question [11]. Expensive, risky, new alternatives or technologies are often associated with high proportions of non-choice, either because respondents are waiting to see if better options arise or because they have decided that they have no interest now or ever [120]. Where status quo options are allowed, they should be defined or respondents asked explicitly how they defined and used status quo in the DCE; otherwise it is difficult to know or understand what the ‘opt-out’ or ‘no choice’ actually means. Opt-out options were used in approximately one-third of studies reviewed, and status quo opt-outs were defined, particularly in studies published in the final 2 years of the review.

4.3 Communication of Risk Attributes

4.3.1 Presentation

The communication of risks and probabilities and the distinction between risks is problematic [121], more so when describing risks with low occurrence probabilities or small changes in risk, typical of healthcare interventions [3]. There was evidence of risks being presented with comparable numerators but varying denominators (denominator neglect), both within and between attributes, which may lead respondents to misinterpreting differences in magnitude between two levels or types of risk. Although there were no examples in the literature of ratio bias, it appears that the recommendations from the risk communication of using a consistent format, including a consistent denominator, for risks and benefits should be reiterated for researchers designing DCEs [4, 5].

A number of studies chose to describe risks qualitatively as an uncertainty, circumventing the problems associated with quantitative presentation of risk, but this may provide a different impression of the changes in risk presented. Qualitative description of risk provides scope for greater flexibility of description but may lead to variability, imprecision and a lack of consistency in how the magnitude of risk is interpreted. The review of risk communication tools suggested this was one of the least effective methods of risk communication. Furthermore, there was inconsistency in the terminology used to describe risks of a similar magnitude when qualitative descriptions were used to support quantitative estimates, and a lack of transparency relating to the likely magnitude of risk when a purely qualitative approach to risk (or uncertainty) presentation was used.

In line with recommendations from the risk communication literature, the identified DCEs were increasing using visual aids (such as icons/dot arrays) to support information communication. It is not clear whether different supplementary aids produce unintended consequences (i.e., demand effects). For example, dramatic, emotional photographs may dominate other information in the DCE [119], and the physical space required to incorporate photographs into attributes may draw the respondents’ attention to that particular attribute.

4.3.2 Framing

One study investigated whether framing of risk as positive (e.g., a cancer detected) or negative (e.g., a cancer missed) found that positive framing can increase preferences for an attribute [45], replicating prior findings outside DCEs [122, 123]. However, as with the use of visual aids, it is unclear whether changes in preferences reflect improvements in the communication of risk. An alternative approach to framing risk is to present both positive and negative outcomes in the form ‘out of 100 operations on people like you, we expect 95 to be successful and 5 to be unsuccessful’ to avoid framing bias [124, 125]; no study reviewed reported utilising this approach. In addition, framing probabilities as relative or absolute risk reductions can affect respondent preferences and choices. Relative risk reductions may look larger than absolute risks, causing respondents to be more willing to choose treatments or avoid risks where information is presented using relative risks compared with absolute risks [4, 122, 126–128]. Reassuringly, risk information in DCEs was mainly described using absolute risk reductions, or both relative and absolute effects to give a complete picture of risk, the convention recommended by CONSORT [129].

Contextual factors, such as the timeframe, can influence the credibility of choices and the degree to which respondents believe the risks apply to them [130]. The timeframe for risks presented were only specified by around a third of studies in this review, although it is possible it was outlined in background materials. Recommendations from the risk communication literature emphasise the importance of specifying and then repeatedly reinforcing the time interval to ensure valid risk communication [131].

Testing the face and construct validity of results from DCEs is an important step in improving the communication of risk. DCE studies that checked the construct validity of
risk preference results tended to be studies conducted in line with other recommendations of good practice from the risk communication and CVM literature, such as supplementing risk data with visual aids. The need for validation of and presentation of risk attributes suggests a role for greater interdisciplinary collaboration with experts from other fields such as psychology when framing and describing attributes. Tests of validity of risk preferences from DCEs associated with risk attributes, by comparing results with those expected on the basis of economic theory, is important for identifying the best ways to communicate risks to patients.

4.4 The Study Sample

Preferences from a range of study samples were included in the DCEs reviewed. It is important to distinguish ex-ante (prior to an event/anticipated) and ex-post (after the event/experienced) preferences. Direct experience of a health condition may affect preferences, for example those who have not experienced a condition or event may overestimate the dreadfulness of bad events (and underestimate the ability to adapt to changes in circumstances) [132], and levels of WTP for a life saved for a curative intervention (ex-post) may exceed those for prevention (ex-ante) [133]. Furthermore, if respondents do not treat a risk as applicable to them, they may use prior beliefs, experiences and the information to subjectively interpret risk information presented [134]. Risks in healthcare often have low occurrence probabilities, which pose a considerable problem, and although healthcare professionals should be experienced and knowledgeable about these risks, they may struggle to use probabilities to judge the level of risk [125]. There are also unique challenges to presenting risk which reflect patient decision making compared with other settings, such as the role of physicians in shared decision making.

Early recommendations of good practice in the CVM literature suggest separating out estimates of WTP by knowledge, relevance and immediacy, and understanding and belief [119]. Whilst the risk communication literature recommends that risk communication strategies may need to be targeted for each setting or audience [4, 124], Table 2 is indicative of a lack of variation in the way risk preferences are assessed in the DCE literature. Careful consideration of the context and consultation with methodologists from other disciplines may be beneficial when designing the optimal format for risk attributes.

4.5 Inclusion of Supplementary Questions

It would seem reasonable to understand the risk preferences of respondents when risk is included in a DCE. Current practice in DCEs seems to assume that respondents’ risk attitudes are neutral, consistent with the cost-effectiveness analysis and decision analysis literature [135, 136]. Risk neutrality assumes people have linear monotonic utility functions in risk; however, it is possible that people may be risk averse or risk prone-seeking. Prospect theory predicts that people may be more risk averse when considering potential losses and risk prone-seeking over gains [123].

Individual determinants of risk preferences, such as risk attitude, numeracy and health literacy, and experience were almost entirely overlooked in either the design or reporting of the identified DCEs. Although risk attitude is known to affect preferences [137], few studies reported attempting to assess respondent risk attitudes. However, there is little evidence that a respondent’s attitude to one type of risk is consistent with their attitude toward bearing health risks, or between different kinds of health risks; for example, one study identified in this review showed that mortality risk tolerance varied across three kinds of mortality risks [138]. Most DCEs asked supplementary questions about respondent characteristics, but questions about knowledge of interventions, risk behaviours or numeracy were less common. The CVM literature recommends stratifying WTP analyses by respondent attitudes and ability and willingness to participate [119].

Numeracy affects how people interpret and understand quantitative risk information [6]. None of the reviewed DCEs reported measuring respondent numeracy and risk literacy. Only one study explicitly discussed methods to detect and overcome problems with numeracy, looking for signs that respondents appeared confused or gave ‘haphazard, inconsistent answers’, and reported strategies to overcome problems with numerical literacy [18].

Although numeracy levels are not reported to influence negative/positive risk framing effects [6], step-by-step risk descriptions that include both positive and negative outcomes to reflect natural decisions and problem solving recommended to overcome framing effects [5] may also aid those with low numeracy [139]. Despite an apparent lack of focus on numerical literacy, the use of natural frequencies and icon arrays was prevalent and these formats are thought to aid low-numeracy groups [86, 124, 140].

4.6 Analysis of Choice Data

DCEs are based on Lancaster’s economic theory of demand and random utility theory (RUT). Incorporating risk introduces another relevant theory, expected utility theory (EUT), concerning decision making under uncertainty. EUT views expected utility to equal the sum of the utilities of the various outcomes multiplied by the probability of each outcome occurring. RUT views the value a person places on a particular service or intervention as
some function of the sum of its separate components or attributes, although RUT allows for incorporation of multiplicative effects through interactions and it is possible to test the assumptions of EUT. Thus, DCEs incorporating risk attributes require additional theoretical links, and DCE designs must support utility functions implied by the theoretical integration.

Most identified DCEs assumed that respondents evaluate risk linearly, incorporating risk attributes as simply as assumed for other attributes. If the indirect utility function is thought to be multiplicative in risk, this must be accounted for at both design and analysis stages. For example, using more general specifications that can capture functional forms for multiple risk attributes and/or implied interactions of these attributes. We found few DCEs that tried to do this (for example, Johnson et al. [79, 138] and van Houtven et al. [79, 138]). One implication of more flexible design strategies is the need to consider larger designs, such as full factorials. Main effects designs are small because they deliberately confound the two-way and higher-order interactions with the main effects. If multiplicative functional forms are introduced, designs that properly separate the main effects from the interactions are needed (making the design larger than a main effects design alone). More general and flexible design strategies are discussed elsewhere [9].

We identified one DCE that reported that the weights respondents place on risks may not be linear but instead fluctuate with the range of probabilities [138], a departure from the assumptions of EUT, and expectations of proportionality of WTP or MRS to the size of risk reduction, but consistent with prospect theory. This work used non-linear weighting techniques to explore the impact of risk-benefit trade-offs under a non-EUT framework.

5 Limitations

The primary limitation of this review is that the methods underlying presentation, format and analysis of risk attributes can only be appraised to the extent that they are reported. Guidelines for authors of peer-reviewed articles restrict manuscript length and may preclude detailed description of the development of surveys, wording and formatting, and other details such as preparing respondents to evaluate risk attributes. It may be that many studies included in this review have performed more rigorous and thorough development of risk attributes than is reported in publications. However, this limitation motivates the need for better reporting of methods, use of (and maintenance of) online appendices if necessary, supporting risk attributes as a key recommendation. More detailed reporting of this process would benefit researchers designing future DCEs and allow the quality of the background and explanatory material to be subject to appraisal to the same extent as the statistical properties of the design and analysis. This recommendation draws strong parallels with recommendations for improved reporting of qualitative methods used in developing and refining attributes for a DCE, which also cites lack of reporting conventions as a barrier to rigorous conduct, limiting the ability of researchers to draw on others' experience, with the consequence of 're-invention' and suppression of debate, which could inspire methodological development [141].

6 Conclusion

This study identified and reviewed a large body of literature that incorporated risk attributes in healthcare DCEs. There was some evidence that some examples of good practice in the presentation of risk, for example the use of absolute risks, and adoption of visual aids to support risk presentation were being adopted. However, little evidence was found that continuing methodological developments and recommendations from other fields about the need for effective risk communication were systematically being applied to DCEs. Specifically, there was limited evidence found about use of training and background material to support risk communication and risk numeracy, and little consideration of the impact of information framing on preferences. Some evidence suggested that different formats were explored to try to help risk understanding and some examples where validity of risk communication was checked using tests of scope scale and linearity. More generally, however, the reviewed DCE studies were not as methodologically thorough in risk presentation as the CVM field, where recommendations on minimal requirements for explaining risk in training materials, consideration of framing of risk in designs, and exploring scope linearity in analyses exist. There seems a clear need for healthcare DCEs to adopt these recommendations. There also appears a real need for those conducting healthcare DCEs to engage with other disciplines to improve the design and presentation of risk attributes; there perhaps is not the amount of attention paid to the specific challenges in developing these attributes as needed. However, to define best practice there is also a need to move beyond statistical testing of the robustness of different methods of presentation towards testing the process of communicating risk in terms of whether people understand risks, which methods change the way in which people interpret risk information, and a framework for identifying which methods are best for the purpose of risk communication. This points to an urgent need for robust multidisciplinary methodological work supported by qualitative and quantitative analysis to inform...
the best way of presenting and analysing risk attributes in DCEs.

Acknowledgments

Professor Katherine Payne was part funded by a Research Council UK (RCUK) Academic Fellowship between September 2007 and September 2012. No other sources of funding were used to conduct this study or prepare this manuscript. Caroline Vass is funded by a National Institute for Health Research (NIHR) School for Primary Care Research Studentship. Mark Harrison, Dan Rigby, Caroline Vass, Terry Flynn, Jordan Lauvere, and Katherine Payne have no conflicts of interest that are directly relevant to the content of this article. The authors are grateful to Mary Ingram, Lawrence Library, Centre for Musculoskeletal Research, University of Manchester, for assistance in developing the electronic search strategies for both reviews, and to Stuart Wright, Manchester Centre for Health Economics, for double screening the abstract lists for inclusion in this review.

Contributions to authorship

Mark Harrison, Katherine Payne, Terry Flynn, Jordan Lauvere and Dan Rigby conceived the idea for this study and produced the protocol which informed the systematic search strategy. Mark Harrison and Katherine Payne applied the inclusion criteria and screening of abstracts, and extracted data from the included studies. Caroline Vass conducted the rapid review. All authors contributed to the writing of the manuscript. Katherine Payne acts as the overall guarantor.

References


68. Schilder LC, Feinberg PM. Alleviating the constant stochastic variance assumption is decision research: theory, measurement, and experimental test. Mark Sci. 2010;29(3):393-412.


Appendix 3.4: List of DCEs with a focus on breast cancer and screening from the systematic review


X16  Howard, K., & Salkeld, G. (2009). Does attribute framing in discrete choice experiments influence willingness to pay? Results from a discrete choice experiment in screening for colorectal cancer. Value in Health, 102(6), 972-980


Appendix 3.5: The design of the DCE used in the pilot study

Figure A1: The Ngene® syntax used for the design of the DCE in the pilot study

Design

;alts = alt1, alt2, alt3
;rows = 40
;block=4
;eff=(mnl,d)

;model:

U(alt1) = s[(n,0.4,0.2)] *saved[3,7,10,14] + r[(n,-0.05,0.02)]*risk[1,5,10,20] + p[(n,-0.5,0.2)]*price[1,2.5,7.5,10] + sr[-0.001]*saved*risk + ps[-0.001]*price*saved + pr[0.001]*price*risk  /
U(alt2) = s*saved + r*risk + p*price + sr*saved*risk + ps*price*saved + pr*price*risk $
Table A1: The Ngene® efficiency measures for the multinomial logit in the pilot study

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>D error</td>
<td>0.00043</td>
<td>0.0009</td>
<td>0.00076</td>
<td>0.000649</td>
<td>0.000352</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A error</td>
<td>0.00906</td>
<td>0.0225</td>
<td>0.027613</td>
<td>0.013891</td>
<td>0.004997</td>
<td>0.222763</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B estimate</td>
<td>41.4266</td>
<td>27.946</td>
<td>14.3455</td>
<td>25.48795</td>
<td>3.449105</td>
<td>78.86291</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S estimate</td>
<td>984.542</td>
<td>3508.0</td>
<td>6327.503</td>
<td>1493.657</td>
<td>633.2704</td>
<td>55365.97</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior</th>
<th>Main Effects</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Detection</td>
<td>Risk</td>
</tr>
<tr>
<td>Fixed prior value</td>
<td>0.4</td>
<td>-0.05</td>
</tr>
<tr>
<td>Sp estimates</td>
<td>0.40695</td>
<td>3.9496</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>58</td>
</tr>
<tr>
<td>Sp t-ratios</td>
<td>3.07242</td>
<td>9862</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>Sb mean estimates</td>
<td>55.5733</td>
<td>28.097</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>Sb mean t-ratios</td>
<td>2.55596</td>
<td>8359</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>88</td>
</tr>
</tbody>
</table>
Table A2: The Ngene®-generated design of the DCE in the pilot study

<table>
<thead>
<tr>
<th>Choice Set</th>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
<th>Block 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alternative</td>
<td>Alternative</td>
<td>Alternative</td>
<td>Alternative</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1 Detection</td>
<td>3</td>
<td>7</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>risk</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Cost</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2 Detection</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>risk</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Cost</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>3 Detection</td>
<td>10</td>
<td>10</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>risk</td>
<td>20</td>
<td>1</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Cost</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>4 Detection</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>risk</td>
<td>1</td>
<td>20</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Cost</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>5 Detection</td>
<td>10</td>
<td>14</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>risk</td>
<td>1</td>
<td>20</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Cost</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>6 Detection</td>
<td>3</td>
<td>10</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>risk</td>
<td>20</td>
<td>5</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Cost</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>7 Detection</td>
<td>3</td>
<td>14</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>risk</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Cost</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>8 Detection</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>risk</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Cost</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>9 Detection</td>
<td>3</td>
<td>14</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>risk</td>
<td>5</td>
<td>20</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Cost</td>
<td>8</td>
<td>10</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>10 Detection</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>risk</td>
<td>10</td>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Cost</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

Detection: 3=3%; 7=7%; 10=10%; 14=14%
Risk: 1=1%; 5=5%; 10=10%; 20=20%
Cost: 1=£100 (£20/screen); 3=£250 (£50/screen); 8=£750 (£150/screen); 10=£1,000 (£200/screen)
### Appendix 3.6: Review of questions used to elicit individual characteristics in commonly-used surveys

<table>
<thead>
<tr>
<th></th>
<th>Census</th>
<th>BHPS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td><strong>Question</strong> What is your date of birth?</td>
<td>Can I just check your age?</td>
</tr>
<tr>
<td><strong>Answer</strong></td>
<td>Free-text/comment</td>
<td>Free-text/comment</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td><strong>Question</strong> Which of these qualifications do you have? Tick every box that applies if you have any of the qualifications listed. If your qualification is not listed, tick the box that contains its nearest equivalent.</td>
<td>Have you attended any education institution full-time since September?</td>
</tr>
<tr>
<td><strong>Answer</strong></td>
<td>• 1-4 O levels…</td>
<td>• Yes</td>
</tr>
<tr>
<td></td>
<td>• NVQ level 1…</td>
<td>• No</td>
</tr>
<tr>
<td></td>
<td>• 5+ O levels…</td>
<td></td>
</tr>
<tr>
<td><strong>Country of birth</strong></td>
<td><strong>Question</strong> What is your country of birth?</td>
<td>Where were you born?</td>
</tr>
<tr>
<td><strong>Answer</strong></td>
<td>• England</td>
<td>Free-text/comment</td>
</tr>
<tr>
<td></td>
<td>• Wales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Scotland</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Wales</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Northern Ireland</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Republic of Ireland</td>
<td></td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td><strong>Question</strong> What is your ethnic group?</td>
<td>To which of these ethnic groups do you consider you belong?</td>
</tr>
<tr>
<td><strong>Answer</strong></td>
<td>White</td>
<td>White</td>
</tr>
<tr>
<td></td>
<td>• British</td>
<td>• British</td>
</tr>
<tr>
<td></td>
<td>• Irish</td>
<td>• Mixed</td>
</tr>
<tr>
<td></td>
<td>• Gypsy/traveller</td>
<td>• White and Black Caribbean</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>• White and Black African</td>
</tr>
<tr>
<td></td>
<td>• White and Black Caribbean</td>
<td>• White and Asian</td>
</tr>
<tr>
<td></td>
<td>• White and Black African</td>
<td>• Asian or Asian British</td>
</tr>
<tr>
<td></td>
<td>• White and Asian</td>
<td>• Indian</td>
</tr>
<tr>
<td></td>
<td>• Other</td>
<td>• Pakistani</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>• Bangladeshi</td>
</tr>
<tr>
<td></td>
<td>• Indian</td>
<td>Black or Black British</td>
</tr>
<tr>
<td></td>
<td>• Pakistan</td>
<td>• Caribbean</td>
</tr>
<tr>
<td></td>
<td>• Bangladeshi</td>
<td>• African</td>
</tr>
<tr>
<td></td>
<td>• Chinese</td>
<td>• Chinese</td>
</tr>
<tr>
<td></td>
<td>• Other</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• African</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Caribbean</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Other</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Arab</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Other</td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td>Question</td>
<td>Census Answer</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td>What is your religion?</td>
<td>● No religion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Christian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Buddhist</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Hindu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Jewish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Muslim</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Sikh</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Other</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Catholic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Presbyterian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Church of Ireland</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Methodist</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Baptist</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Presbyterian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Brethren</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Protestant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Buddhist</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Hindu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Jewish</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Muslim</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Sikh</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● No religion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Employment status</th>
<th>Question</th>
<th>Census Answer</th>
<th>BHPS Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Last week, were you: tick all that apply, include any paid work.</td>
<td>● Working as an employee?</td>
<td>● Last week were you in paid employment at all, including being away temporarily from a job you would normally have been doing?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● On a government sponsored training scheme?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Self-employed or freelance?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Working paid or unpaid or your family’s business?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Away from work ill, on maternity leave, on holiday or temporarily laid off?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Doing any other kind off paid work?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>● None of the above</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income</th>
<th>Question</th>
<th>N/A</th>
<th>Census Answer</th>
<th>BHPS Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>On average, what was your MONTHLY income from this job/business over the last 12 months?</td>
<td>Free-text/comment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3.7: Confirmation letter of University Research Ethics approval

Miss. Vass
Centre for Health Economics
Institute for Population Health
3rd October 2013

Dear Miss. Vass

Research Ethics Committee 3

Vass, Payne, Rigby, Campbell: Using discrete choice experiments to value risks and benefits in primary care (ref 13178)

I write to confirm that the Chair is now satisfied that you have addressed the concerns of the Ethics Committee of the 18th of September 2013 and has therefore given the above research project a favourable ethical opinion.

This approval is effective for a period of five years and if the project continues beyond that period it must be submitted for review. It is the Committee’s practice to warn investigators that they should not depart from the agreed protocol without seeking the approval of the Committee, as any significant deviation could invalidate the insurance arrangements and constitute research misconduct. We also ask that any information sheet should carry a University logo or other indication of where it came from, and that, in accordance with University policy, any data carrying personal identifiers must be encrypted when not held on a university computer or kept as a hard copy in a location which is accessible only to those involved with the research.

Finally, I would be grateful if you could complete and return the attached form at the end of the project or by September 2014.

We hope the research goes well.

Yours sincerely,

Adrian Jarvis
Ethics Committee 3 Secretary
Appendix 3.8: The DCE survey used for the pilot study

Which breast screening programme would you prefer?

A survey to find out how you balance the risks and benefits in breast screening.

Click the button below to enter the survey.

Manchester Centre for Health Economics
We are interested in how women understand the different risks and benefits of a national breast screening programme.

You have been invited to take part in this survey because, as a woman aged between 18 and 70, you may have considered, or may consider in the future, taking part in a breast screening programme.

The survey will take around 20 minutes to complete.

I have read the participant information above and wish to participate in this study.

I do NOT want to continue

If you have any questions, please contact:

Caroline Vass
Manchester Centre for Health Economics
John MacFarlane Building
The University of Manchester
Oxford Road
Manchester
M13 9PL
Caroline.vass@postgrad.manchester.ac.uk
Tel: 077 1675 1892
What age category do you fall into?

- Background A: 18-24
- Background B: 25-34
- Background C: 35-44
- Background D: 45-49
- Background E: 50+
What is breast cancer and what is breast screening?

Breast cancer is a group of cells which can grow in an abnormal way and form into a lump (known as a cancer or tumour). As the lump grows, cancerous cells can spread to other parts of the body and this can be very serious.

Breast screening is a method of detecting breast cancer at a very early stage. Screening involves an x-ray of each breast (a mammogram) which is taken while carefully compressing the breast. Screening for breast cancer has been shown to be safe although most women find it a bit uncomfortable, and a few find it painful.

The x-ray can detect small changes in breast tissue which may indicate a very early stage cancer. These cancers are generally too small to be detected by touch. If there is something unusual on the x-ray then the woman will be asked to return for further tests to check whether this is a cancer.
As mentioned previously, the x-ray images are not always clear and sometimes further tests are needed to confirm if cancerous cells are present. These tests usually require a small sample of tissue (a biopsy) to be taken from the breast by a doctor, using a long needle, under local anaesthetic. Even when there is a tumour, sometimes these small lumps are treated with surgery and chemotherapy when really they would not have grown into something dangerous. It is impossible for doctors to tell which tumours will become life-threatening so everything identified in screening is treated.

This means that screening for breast cancer can lead to further tests and procedures which may have been unnecessary. There will be a delay between the tests and hearing the results which might be stressful for you. Also, the procedures, such as the biopsy, can be uncomfortable or painful.
This figure shows the different options that could happen in screening.

- Woman aged 50-70 invited for screening
- Mammography (X-ray of breasts)
- Biopsy required?
- Treatment offered?
- Outcome

Positive Result
- NO CANCER
- YES: Cancerous
  - NO CANCER
  - CANCER

Negative Result
- NO CANCER
- YES: Cancerous
  - NO CANCER
  - CANCER

2 weeks after mammogram to results.
In the UK, women over the age of 50 are offered screening up until the age of 70. Although the service is free, women who participate in screening incur some cost. For example, they may have to take some time off work, travel to the centre for breast screening, pay for parking and transport, or maybe pay for childcare.

There are **benefits** to breast screening for cancer which means tumours are picked up earlier and the long-term health outcomes of women are improved. However, there are **risks** because it is inconvenient and can mean women undergo procedures unnecessarily.

In this survey we are interested in understanding how women balance these benefits and risks. To do that, we would now like you to complete this survey where you will be shown different screening programmes. You will be asked to choose your preferred programme from a set of possible screening programmes.

**There are no wrong or right answers; we are just interested in your views**
**Risk** is a term used to explain the chance that something bad might happen. In breast cancer screening, there is a chance that women might have unnecessary treatment after their mammogram. This will not happen to all women, but there is a risk it could happen to some women.

The women shaped figures represent the total number of women having screening. The purple coloured figure shows the one woman who went for screening and had unnecessary treatment. This means 1 out of every 100 women screened was treated when she didn’t need to be, but 99 women were treated appropriately.

---

**Risk** is a term used to explain the chance that something bad might happen. In breast cancer screening, there is a chance that women might have unnecessary treatment after their mammogram. This will not happen to all women, but there is a risk it could happen to some women.

A risk of 1% means that 1 out of every 100 women screened was treated when she didn’t need to be, but 99 women were treated appropriately.
This grid shows that 5 out of every 100 women screened was treated when she didn't need to be, but 95 women were treated appropriately.

OR

Similarly, a risk of 5% means that 5 out of every 100 women screened was treated when she didn't need to be, but 95 women were treated appropriately.
You are going to be asked to choose between screening programmes which differ in terms of:

**Women who will have cancers detected by screening**
This describes the number of women who will have a cancer identified by a mammogram (x-ray of the breast). This describes the number of women, who go for screening over a period of 20 years, that are identified as having a cancer.

**Women who will have unnecessary treatment**
This shows the number of women requiring a procedure when the lump was not cancerous or would not have caused any harmful symptoms in their lifetime. The unnecessary treatment includes biopsies and possibly further surgery and chemotherapy.

**Out-of-pocket cost**
This describes the cost to you of participating in a screening programme. There are no plans to introduce charges for screening with the NHS. This cost reflects your out-of-pocket expenses incurred when participating in the screening programme. For example, transport costs, car parking, child care, time off work. In the UK, women will be screened 5 times over a 20 year period (a lifetime programme). We’ll show the cost of the programmes over your life time and what that works amounts to per screen.

Each screening programme we show you will be described in terms of these three characteristics.

Each characteristic, however, has four different levels that will change.
Women who will have cancers detected by screening

3% means that if 100 women are screened for 20 years, then 3 women will be identified as having cancer.

7% means that if 100 women are screened for 20 years, then 7 women will be identified as having cancer.

10% means that if 100 women are screened for 20 years, then 10 women will be identified as having cancer.

14% means that if 100 women are screened for 20 years, then 14 women will be identified as having cancer.
Women who will have an unnecessary treatment

1% means that if 100 women are screened for 20 years, then 1 woman will be treated unnecessarily.

5% means that if 100 women are screened for 20 years, then 5 women will be treated unnecessarily.

10% means that if 100 women are screened for 20 years, then 10 women will be treated unnecessarily.

20% means that if 100 women are screened for 20 years, then 20 women will be treated unnecessarily.
Out-of-pocket cost to you of screening programme per year

- £20 (or £100 over your lifetime)
- £50 (or £250 over your lifetime)
- £150 (or £750 over your lifetime)
- £200 (or £1,000 over your lifetime)
Below is an example of one possible breast screening programme.

<table>
<thead>
<tr>
<th>Programme A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers</td>
<td>10%</td>
</tr>
<tr>
<td>detected by screening</td>
<td></td>
</tr>
<tr>
<td>Women who will have an</td>
<td>20%</td>
</tr>
<tr>
<td>unnecessary follow-up</td>
<td></td>
</tr>
<tr>
<td>Cost to you</td>
<td></td>
</tr>
<tr>
<td>£20 per screen</td>
<td></td>
</tr>
<tr>
<td>(£100 over your lifetime)</td>
<td></td>
</tr>
</tbody>
</table>

OR

Below is an example of one possible breast screening programme.

<table>
<thead>
<tr>
<th>Programme A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers</td>
<td>10%</td>
</tr>
<tr>
<td>detected by screening</td>
<td></td>
</tr>
<tr>
<td>Women who will have an</td>
<td>20%</td>
</tr>
<tr>
<td>unnecessary follow-up</td>
<td></td>
</tr>
<tr>
<td>Cost to you</td>
<td></td>
</tr>
<tr>
<td>£20 per screen</td>
<td></td>
</tr>
<tr>
<td>(£100 over your lifetime)</td>
<td></td>
</tr>
</tbody>
</table>
Below is an example of the type of question we will ask you.

If these were your only options, which would you choose?

<table>
<thead>
<tr>
<th>Women who will have cancers detected by screening</th>
<th>Programme A</th>
<th>Programme B</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td>Women who will have an unnecessary follow-up</td>
<td>20%</td>
<td>1%</td>
</tr>
<tr>
<td>Cost to you (£20 per screen) ($100 over your lifetime)</td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
</tbody>
</table>

You simply select the programme which you prefer.

OR

Below is an example of the type of question we will ask you.

If these were your only options, which would you choose?

<table>
<thead>
<tr>
<th>Women who will have cancers detected by screening</th>
<th>Programme A</th>
<th>Programme B</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>20%</td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>1%</td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td>Cost to you (£20 per screen) ($100 over your lifetime)</td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
</tbody>
</table>

You simply select the programme which you prefer.
Thinking back over what you just read, is there anything about the breast screening information you've been given that seems unclear? (e.g. words that are confusing or things which could be phrased better)

- Feedback B3a: No, I understand everything
- Feedback B3b: There are some things that aren't clear or well explained (please comment in the box below)

Is there any more information or further explanation you would like?
Please press the “next” arrow below to begin the survey.
If these were your only options, which would you choose?
Choose by clicking one of the buttons below:

(1 of 10 choice questions)

<table>
<thead>
<tr>
<th>Programme A</th>
<th>Programme B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Women who will have cancers detected by screening</strong></td>
<td>10%</td>
</tr>
<tr>
<td><strong>Women who will have an unnecessary follow-up</strong></td>
<td>20%</td>
</tr>
<tr>
<td><strong>Cost to you</strong></td>
<td>£20 per screen (£100 over your lifetime)</td>
</tr>
</tbody>
</table>

NONE: I wouldn't choose any of these.

OR

If these were your only options, which would you choose?
Choose by clicking one of the buttons below:

(1 of 10 choice questions)

<table>
<thead>
<tr>
<th>Programme A</th>
<th>Programme B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Women who will have cancers detected by screening</strong></td>
<td>10%</td>
</tr>
<tr>
<td><strong>Women who will have an unnecessary follow-up</strong></td>
<td>20%</td>
</tr>
<tr>
<td><strong>Cost to you</strong></td>
<td>£20 per screen (£100 over your lifetime)</td>
</tr>
</tbody>
</table>

NONE: I wouldn't choose any of these.
Thank you for providing your choices. You will now be asked some questions about the survey and a few questions about yourself.
How confident are you that you would make the same choices if faced with the situations in real-life?

- FeedbackA: Very confident I would make the same choices
- FeedbackB: Quite confident I would make the same choices
- FeedbackC: Not confident I would make the same choices
- FeedbackD: Confident I would make different choices

Please explain your answer:
On a scale of 1-5, how easy or difficult did you find making choices between the alternatives?

<table>
<thead>
<tr>
<th></th>
<th>Very easy</th>
<th>Neither difficult nor easy</th>
<th>Very difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FeedbackB_r1=1</td>
<td>FeedbackB_r1=2</td>
<td>FeedbackB_r1=5</td>
</tr>
<tr>
<td>2</td>
<td>FeedbackB_r1=3</td>
<td>FeedbackB_r1=4</td>
<td>FeedbackB_r1=5</td>
</tr>
<tr>
<td>3</td>
<td>FeedbackB_r1=5</td>
<td>FeedbackB_r1=5</td>
<td>FeedbackB_r1=5</td>
</tr>
</tbody>
</table>

Please explain your answer

---

FeedbackB_r1=1
FeedbackB_r1=2
FeedbackB_r1=3
FeedbackB_r1=4
FeedbackB_r1=5

0.00%

---

MANCHESTER CENTRE FOR HEALTH ECONOMICS

National Institute for Health Research
School for Public Health Research
On a scale of 1-5, how easy or difficult did you find the survey to understand?

<table>
<thead>
<tr>
<th>1 (Very easy)</th>
<th>2</th>
<th>3 (Neither difficult nor easy)</th>
<th>4</th>
<th>5 (Very difficult)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

Was there anything in the survey that was particularly unclear?

- [ ] No, I understood everything
- [ ] [ ] There were some things that weren’t clear or well explained

(please comment in the box below)

Is there anything else about the survey we could change to make it easier to understand (e.g. phrasing, explanations)?

[ ]
Did you find yourself making choices based on one or two characteristics of the screening programme?

Yes, I based choices on one or two characteristics
No, I considered ALL characteristics

Would you say you considered the characteristic women who have a cancer detected?
Please select

Would you say you considered the characteristic risk of unnecessary treatment?
Please select

Would you say you considered the characteristic cost?
Please select

If you based your choices on one or two characteristics, why was this? (Please tick all that apply):

There was just too many characteristics on which to make a decision
The other characteristic(s) were too complicated or it was unclear what it meant
Other reason (please explain)
What is your occupational status?

- Employed full-time
- Employed part-time
- Self-employed
- Unemployed
- Retired
- Looking after a home/family
- Student
- Freelance or temping
- Long-term sickness
- Temporarily laid off

What is your occupation?
(Or previous occupation if retired?)
What is the highest level of education you have obtained?

- No formal qualifications
- 1-4 O-levels/GCSEs
- 5+ O-levels/GCSEs
- NVQs
- A-levels/AS-levels
- Undergraduate degree
- Master's degree
- PhD
- Other formal qualification
What is your religious group?

- No religion
- Christian
- Buddhist
- Jewish
- Hindu
- Muslim
- Sikh
- Other (please specify)
Do you have any children?

- Background=1 Yes
- Background=2 No
Do you have any daughters?
- Yes
- No
Have you ever had breast screening (a mammogram)?

- Yes
- No
- Don't know

Have you received follow-up for breast cancer?

- Yes
- No
- Don't know
Has anyone in your family (blood relative or otherwise) had breast cancer?

- Yes
- No
- Don't know

Do you know of anybody else (e.g., friends or colleagues) who has had breast cancer?

- Yes
- No
- Don't know
How would you describe your current risk of breast cancer?
- High or very high
- Quite high/above average
- Average
- Slight or low
- No idea

How would you describe your future risk of breast cancer?
- High or very high
- Quite high/above average
- Average
- Slight or low
- No idea

To what extent are you concerned about your own risk of breast cancer?
- Not at all
- A little
- Quite a lot
- Very much
- No idea

Is there anything else you would like to tell us about your risk of breast cancer?
On a scale of 1-5, how much do you agree with the following statements?

"I feel life is very much out of our hands, and fate determines everything"

1 (Strongly disagree) 2 (Disagree) 3 (Neither agree nor disagree) 4 (Agree) 5 (Strongly agree)

"I get a thrill when I gamble, even if it’s just the lottery"

1 (Strongly disagree) 2 (Disagree) 3 (Neither agree nor disagree) 4 (Agree) 5 (Strongly agree)

"Even if something is unlikely, I feel it will probably happen to me"

1 (Strongly disagree) 2 (Disagree) 3 (Neither agree nor disagree) 4 (Agree) 5 (Strongly agree)

"I have an idea about whether or not I’ll develop breast cancer"

1 (Strongly disagree) 2 (Disagree) 3 (Neither agree nor disagree) 4 (Agree) 5 (Strongly agree)
This next section will ask you some questions to understand how familiar you are with probabilities.

If you are unsure of the answer, you can skip the question.

Imagine I flip a coin 1,000 times. What is your best guess about how many times the coin would fall heads up in 1,000 flips?

If there was a lottery where the chance of winning £10 is 1%, if 1,000 people bought a ticket, how many people would you expect to win?

In another lottery, the chance of winning a car is 1 in 1,000. What percentage of tickets in the lottery will win a car?
Did any of the questions in the second half of the survey (about you, risk or breast cancer) seem confusing or unclear? (You can use the arrows to look back)

Can you think of any other questions which might help us understand how you made your choices in the choice task?

If you have any further comments or suggestions about the survey or the issues it raised then please let us know using the space here.
Thank you for completing this survey.

If you have been affected by any of the issues raised in this survey, the following organisations may be of assistance:

- National Hereditary Breast Cancer Helpline on 01629 813000
- Breast Cancer Care on 0800 668 6600

If you have any questions or concerns please contact Caroline Vass on 077 1675 1262 or caroline.vass@postgrad.manchester.ac.uk.

Manchester Centre for Health Economics, Jean McFarlane Building, University of Manchester, Oxford Road, Manchester M13 9PL
Appendix 3.9: Online advertisement for study participants

UNIVERSITY OF MANCHESTER

My Manchester

You are here: My Manchester > Research volunteering > Opportunity

Research volunteering
- Volunteering Opportunity Submission
- Guidelines for recruiting volunteers
- Removing your request
- Contact us
- Join mailing list

Share this page:

UNDERSTANDING RISK IN HEALTHCARE: AN INTERVIEW STUDY

DESCRIPTION

We make choices that involve risk all the time. How we travel, what we choose to eat, all involves balancing the risks and benefits. We are interested in how you balance risks and benefits in healthcare, particularly breast cancer screening.

We are looking for female volunteers aged 18-70 to take part in an interview. The interview will involve the use of colourful images so volunteers with red-green colour blindness are not eligible to partake in this study. The interview should take no more than 45 minutes and can take place on the university campus or at a location convenient for you.

If you would like to find out more information please contact Caroline Vass.

Name of Research Ethics Committee: University Research Ethics Committee 3
University Ethics Committee number and/or NHS Reference number (if applicable): AJ1ethos/1809/13

CONTACT DETAILS

- caroline.vass@postgrad.manchester.ac.uk or 0771 675 1262
Appendix 3.10: Preliminary interview schedule for pilot study

Introduction:
My name is Caroline Vass and I’m a PhD student at the Manchester Centre for Health Economics. My PhD is about how people make a balance between risks and benefits when making choices in health care. I am particularly interested in how we can use surveys to understand their views on risks and benefits. This interview is about how you balance risks and benefits in health care. I will ask you to complete a questionnaire and to say out-loud how you have come to your answers.
The interview should take about 45 minutes, and there will be imaginary examples looking at breast cancer screening programmes.
Have you read the information sheet? Please complete this consent form. So, are you ok to continue?

Topic 1: The DCE
Here is a survey made up of questions about imaginary breast cancer screening programmes. Take a few minutes to read through the information and definitions at the front of the survey. Do you have any questions at this point?

Instruction 1: As you complete the survey I want you to talk aloud by trying to say out-loud what you’re thinking as you’re completing this questionnaire.

Don’t worry about making sense or talking to me, I am merely going to listen to what you have to say. If you go a bit quiet, I might ask some questions to get you talking again. Is that ok?
Probes:
- What are you thinking now?
- Why did you choose that one?
- Would you choose that if I wasn’t here?
- Are you considering all of the information presented?

[If respondent is choosing A’s or B’s, probe why…]
[If respondent fails internal tests [QUESTION 6], probe why…]

Topic 2: Ease of Completion
Question 1: What did you think about the questions?

Question 2: Did you find yourself concentrating on a particular characteristic or do you think you weighed them up evenly?

Question 3: Would you change any of your answers?

Question 4: Do you think you are satisfied with your choices?

Topic 3: Further questions
These questions are just so I can understand if different people see things differently.

Question 1: What is your current employment status? What kind of work do / did you do? Was/Is that full or part-time? Do you have to use numbers in this job?

Question 2: Do you play games that involve a gamble? If not clear, suggest betting on horses or bingo. Where do you play <insert game>? Is that online?

Topic 4: Final thoughts
Do you have any other feedback or thoughts?
Thank you for completing this interview.
Appendix 3.11: Information sheet for interviewees for the qualitative interview study

Understanding risk in healthcare: an interview study

You are being invited to take part in a research study that aims to understand how people complete surveys. The example we will use is choosing a screening programme for breast cancer. Before you decide whether to take part in this study it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully. Please ask if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

What is the aim of the research?
This study wants to understand how people make choices about healthcare. The aim is to explore how people balance the risks and benefits when making a decision about breast cancer screening.

Why have I been chosen?
This study aims to explore the views of women. Any female member of the public who is able to complete an interview in English is able to take part in this study.

What would I be asked to do if I took part?
You will be invited to complete an interview at a location convenient to you or on University of Manchester premises if you prefer. The interview should take no more than 45 minutes. During the interview I will ask you to complete a questionnaire with different imaginary breast cancer screening programmes and you will be asked which you prefer. It is important to remember that there are no right or wrong answers; we are just trying to explore your views. I will also ask your permission to record the interview so that I do not have to take notes to remember what was said.
What happens to the data collected?
The interview will be recorded and then typed up into a written record of what was said. The answers to the survey will be used to help us understand how breast cancer screening programmes can be better provided. Your name will not be recorded and each interview will be given an anonymous identification number. We might want to use some of the direct quotes from the interview in a document and/or research paper summarising the findings but your name will not be recorded with these quotes.

How is confidentiality maintained?
The law called the Data Protection Act (1998) tells us how to keep your information secure. If you provide personal information (name, address, contact details) it will remain confidential and we will not give your details to anyone else.

Will I be paid for participating in the research?
We are not able to pay you for taking part in this study. If you decide to travel to the University for the interview then we can reimburse your travel expenses in line with university policy. To thank you for your time, you will receive an Amazon voucher for participating in this research.

Where will the research be conducted?
The interview will take place at a time and location convenient for you where it is possible to record what was said privately and clearly. If you prefer, then you are free to travel to the University for the interview. The location of the interview is your choice.

Will the outcomes of the research be published?
The main outcome of this research will be a PhD thesis for Caroline Vass, a student based in the Manchester Centre for Health Economics at The University of Manchester. In addition, we may want to report the findings at conferences or in a published journal article.

My question is not covered here, who do I contact for more information?
Please contact Caroline Vass (PhD Student in the Centre for Health Economics) on 077 1675 1252 or email caroline.vass@postgrad.manchester.ac.uk.

Thank you for reading this information sheet
Please keep this information sheet for your records
Appendix 3.12: Consent form for interviewees for the qualitative interview study

Consent form
Understanding risk in healthcare

Please tick box

1. I confirm that I have read the attached information sheet on the above project and have had the opportunity to consider the information and ask questions and have these answered satisfactorily.

2. I understand that my participation in the study is voluntary and that I am free to withdraw at any time without giving a reason.

3. I understand that the interviews will be digitally voice-recorded and notes may be taken.

4. I agree to the use of anonymous quotes in academic articles.

I agree to take part in the above project

____________________________  __________________  __________________
Name of participant          Date                     Signature

____________________________  __________________  __________________
Name of person taking consent Date                     Signature
Appendix 3.13: NHS Choices video approval

RE: Feedback - Content - 177133 JA
NHS Choices Service Desk [servicedesk@nhschoices.nhs.uk]
Sent: 08 May 2014 11:52
To: Caroline Vass

Dear Caroline,
Thank you for contacting the NHS Choices Service Desk.
Please see the response below from our Video Content team.

"Hi Caroline
We would be happy for you to use the video. You can either link to the video, or embed the video in your webpage via the embed code on the video player."

Please get back to us if you have any other queries.
Kind Regards,
James
The NHS Choices Service Desk

-----Original Message-----
From: NHS Choices Service Desk
Sent: 04 April 2014 13:03
To: 'Caroline Vass'
Subject: RE: Feedback - Content - 177133 JA

Dear Caroline,
Thank you for contacting the NHS Choices Service Desk.
We have now assigned your query to the Subject Matter Expert (SME) within NHS Choices. A member of our service desk will be in touch as soon as we receive a response.
If you require an update at any time, please feel free to respond to this email quoting the reference 177133 in the subject line.
Kind Regards,
James
The NHS Choices Service Desk
Appendix 3.14: Pilot quantitative results of the DCE from the internet panel

Table A1: Response statistics

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Icon arrays</th>
<th>Percentages only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed to termination</td>
<td>56</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Completed choice questions</td>
<td>62</td>
<td>31</td>
<td>31</td>
</tr>
</tbody>
</table>

Table A2: Effect of attributes on choice probabilities and interactions (conditional logit)

<table>
<thead>
<tr>
<th></th>
<th>No Interactions</th>
<th>Two-way Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect</td>
<td>0.142*** (0.02)</td>
<td>0.145*** (0.02)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.035*** (0.01)</td>
<td>-0.011 (0.01)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.001*** (0.00)</td>
<td>-0.001*** (0.00)</td>
</tr>
<tr>
<td>Detect*risk</td>
<td></td>
<td>-0.003 (0.00)</td>
</tr>
<tr>
<td>Detect*cost</td>
<td></td>
<td>0.000 (0.00)</td>
</tr>
<tr>
<td>Risk*cost</td>
<td></td>
<td>-0.011 (0.01)</td>
</tr>
<tr>
<td>Observations</td>
<td>1860</td>
<td>1860</td>
</tr>
</tbody>
</table>

* p<0.05  ** p<0.01  *** p<0.001; Standard errors in parentheses

Table A3: Effect of attributes on choice probabilities by risk communication format (conditional logit)

<table>
<thead>
<tr>
<th></th>
<th>Percentage participants</th>
<th>Icon array participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect</td>
<td>0.107*** (0.03)</td>
<td>0.179*** (0.03)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.013 (0.01)</td>
<td>-0.060*** (0.01)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.001* (0.00)</td>
<td>-0.001** (0.00)</td>
</tr>
<tr>
<td>Observations</td>
<td>930</td>
<td>930</td>
</tr>
</tbody>
</table>

* p<0.05  ** p<0.01  *** p<0.001; Standard errors in parentheses

Table A4: Effect of attributes on choice probabilities with interaction of risk communication format (conditional logit)

<table>
<thead>
<tr>
<th></th>
<th>Detect</th>
<th>Risk</th>
<th>Cost</th>
<th>IAP*detect</th>
<th>IAP*risk</th>
<th>IAP*cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.107*** (0.03)</td>
<td>-0.013 (0.01)</td>
<td>-0.001* (0.00)</td>
<td>0.073 (0.04)</td>
<td>-0.047** (0.01)</td>
<td>-0.000 (0.00)</td>
</tr>
<tr>
<td>Observations</td>
<td>1860</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* p<0.05  ** p<0.01  *** p<0.001; Standard errors in parentheses
Table A5: Effect of attributes on choice probabilities and testing for significance of the scale parameter (heteroskedastic conditional logit)

<table>
<thead>
<tr>
<th></th>
<th>Pooled model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect</td>
<td>0.107*** (0.03)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.013 (0.01)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.070* (0.03)</td>
</tr>
<tr>
<td>IAP*detect</td>
<td>0.016 (0.05)</td>
</tr>
<tr>
<td>IAP*risk</td>
<td>-0.028 (0.02)</td>
</tr>
<tr>
<td>Observations</td>
<td>1860</td>
</tr>
<tr>
<td>Scale term: IAP</td>
<td>0.380 (0.573)</td>
</tr>
</tbody>
</table>

* p<0.05  ** p<0.01  *** p<0.001; Standard errors in parentheses

Table A6: Summary of internal validity tests

<table>
<thead>
<tr>
<th>Risk version</th>
<th>Mean</th>
<th>Std. Dev</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentages only</td>
<td>.19354839</td>
<td>.39529157</td>
<td>930</td>
</tr>
<tr>
<td>Icon arrays</td>
<td>.03225806</td>
<td>.17677976</td>
<td>930</td>
</tr>
<tr>
<td>Total</td>
<td>.11290323</td>
<td>.31655957</td>
<td>1860</td>
</tr>
</tbody>
</table>

Table A7: Willingness-to-Pay estimates (conditional logit)

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Icon Arrays</th>
<th>Percentages Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk</td>
<td>WTP</td>
<td>Detect</td>
<td>Risk</td>
</tr>
<tr>
<td></td>
<td>-40.95747</td>
<td>165.60727</td>
<td>-58.776433</td>
</tr>
<tr>
<td>LL</td>
<td>-63.490184</td>
<td>108.93042</td>
<td>-95.267242</td>
</tr>
<tr>
<td>UL</td>
<td>-18.424909</td>
<td>222.28411</td>
<td>-22.285624</td>
</tr>
</tbody>
</table>

Table A8: Effect of attributes on choice probabilities by risk communication method (mixed logit)

<table>
<thead>
<tr>
<th></th>
<th>Cost</th>
<th>Risk</th>
<th>Detect</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage participants</td>
<td>Icon array participants</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.001*</td>
<td>-0.052</td>
<td>0.255*</td>
<td>0.193***</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.04)</td>
<td>(0.10)</td>
<td>(0.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.561***</td>
<td>0.173***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.13)</td>
<td>(0.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.457***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>930</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>930</td>
</tr>
</tbody>
</table>

* p<0.05  ** p<0.01  *** p<0.001; Standard errors in parentheses

Table A9: Willingness-to-Pay estimates (mixed logit)

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Icon Arrays</th>
<th>Percentages Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk</td>
<td>WTP</td>
<td>Detect</td>
<td>Risk</td>
</tr>
<tr>
<td></td>
<td>-67.52141</td>
<td>252.98664</td>
<td>-68.411058</td>
</tr>
<tr>
<td>LL</td>
<td>-114.29974</td>
<td>128.84089</td>
<td>-119.31961</td>
</tr>
<tr>
<td>UL</td>
<td>-20743079</td>
<td>377.13238</td>
<td>-17.502507</td>
</tr>
</tbody>
</table>
## Appendix 3.15: The design of the DCE used in the final empirical studies

### Table A1: Ngene® efficiency measure for the multinomial logit in the final empirical studies

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>D error</td>
<td>0.000934</td>
<td>0.016483</td>
<td>0.02654</td>
<td>0.007882</td>
<td>0.001087</td>
<td>0.27508</td>
</tr>
<tr>
<td>A error</td>
<td>0.012077</td>
<td>0.609606</td>
<td>2.254365</td>
<td>0.12509</td>
<td>0.011562</td>
<td>29.740286</td>
</tr>
<tr>
<td>B estimate</td>
<td>77.212361</td>
<td>27.702787</td>
<td>14.063811</td>
<td>24.75517</td>
<td>2.212302</td>
<td>84.387436</td>
</tr>
<tr>
<td>S estimate</td>
<td>173.168589</td>
<td>48301.84474</td>
<td>358251.9342</td>
<td>557.376104</td>
<td>14.588552</td>
<td>4479484.918</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior</th>
<th>Detection</th>
<th>Risk</th>
<th>Cost</th>
<th>Detection* Risk</th>
<th>Cost* Detection</th>
<th>Detection* Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed prior value</td>
<td>0.1</td>
<td>-0.06</td>
<td>-0.14</td>
<td>-0.004</td>
<td>-0.015</td>
<td>-0.002</td>
</tr>
<tr>
<td>Sp estimates</td>
<td>7.557505</td>
<td>12.976973</td>
<td>7.845484</td>
<td>20.841688</td>
<td>5.689372</td>
<td>173.168589</td>
</tr>
<tr>
<td>Sp t-ratios</td>
<td>0.712963</td>
<td>0.544088</td>
<td>0.699755</td>
<td>0.429328</td>
<td>0.82172</td>
<td>0.148943</td>
</tr>
<tr>
<td>Sb mean estimates</td>
<td>826.020716</td>
<td>482.413463</td>
<td>20722.56166</td>
<td>2727.330143</td>
<td>23176.52967</td>
<td>1439.73165</td>
</tr>
<tr>
<td>Sb mean t-ratios</td>
<td>0.484684</td>
<td>0.22619</td>
<td>0.307684</td>
<td>0.669557</td>
<td>0.486714</td>
<td>0.60736</td>
</tr>
</tbody>
</table>
Table A2: The Ngene®-generated design of the DCE used in the final study

<table>
<thead>
<tr>
<th>Choice Set</th>
<th>Block 1</th>
<th>Block 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alternative</td>
<td>Alternative</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>Detection</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Detection</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Detection</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Detection</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>Detection</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>8</td>
</tr>
<tr>
<td>6*</td>
<td>Detection</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Detection</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>Detection</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Detection</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>Detection</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>Detection</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>risk</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>10</td>
</tr>
</tbody>
</table>

*Detection: 3=3%; 7=7%; 10=10%; 14=14%  
Risk: 0=0%; 1=1%; 5=5%; 10=10%; 20=20%  
Cost: 1=£100 (£20/screen); 3=£250 (£50/screen); 8=£750 (£150/screen); 10=£1,000 (£200/screen)  
*Internal validity test
Appendix 3.16: The DCE survey used for the final empirical studies

Which breast screening programme would you prefer?

A survey to find out how you balance the risks and benefits in breast screening.
Click the button below to enter the survey.

Manchester Centre for Health Economics
We are interested in how women understand the different risks and benefits of a national breast screening programme.

You have been invited to take part in this survey because, as a woman aged between 18 and 70, you may have considered, or may consider in the future, taking part in a breast screening programme.

The law called the Data Protection Act (1998) tells us how to keep your information secure. If you provide personal information (name, address, contact details) it will remain confidential and we will not give your details to anyone else.

The main outcome of this research will be a PhD thesis for Caroline Vass, a student based in the Manchester Centre for Health Economics at The University of Manchester. In addition, we may want to report the findings at conferences or in a published journal article.

The survey will take around 20 minutes to complete.

[Consent] I have read the participant information above and wish to participate in this study.

[Consent] I do NOT want to continue

Your answers to this survey will remain strictly confidential. Unlike other surveys you may have answered, the results will only be used by The University of Manchester for a project supervised by the NHS.

No results or data will ever be passed to a private company or organisation.

If you have any questions, please contact:

Caroline Vass  
Manchester Centre for Health Economics  
Jenkin McFarlane Building  
The University of Manchester  
Oxford Road  
Manchester  
M13 9PT  
Caroline.vass@postgrad.manchester.ac.uk  
Tel. 077 1875 1292
What age category do you fall into?

- 18-24
- 25-34
- 35-44
- 45-49
- 50+
What is breast cancer and what is breast screening?

Breast cancer is a group of cells which can grow in an abnormal way and form into a lump (known as a cancer or tumour). As the lump grows, cancerous cells can spread to other parts of the body and this can be very serious.

Breast screening is a method of detecting breast cancer at a very early stage. Screening involves an x-ray of each breast (a mammogram) which is taken while carefully compressing the breast. Screening for breast cancer has been shown to be safe although most women find it a bit uncomfortable, and a few find it painful.

The x-ray can detect small changes in breast tissue which may indicate a very early stage cancer. These cancers are generally too small to be detected by touch. If there is something unusual on the x-ray then the woman will be asked to return for further tests to check whether this is a cancer.
INTRO

As mentioned previously, the x-ray images are not always clear and sometimes further tests are needed to confirm if cancerous cells are present. These tests usually require a small sample of tissue (a biopsy) to be taken from the breast by a doctor, using a long needle, under local anaesthetic. Even when there is a tumour, sometimes these small lumps are treated with surgery and chemotherapy when really they would not have grown into something dangerous. It is impossible for doctors to tell which tumours will become life-threatening so everything identified in screening is treated.

This means that screening for breast cancer can lead to further tests and procedures which may have been unnecessary. There will be a delay between the tests and hearing the results which might be stressful for you. Also, the procedures, such as the biopsy, can be uncomfortable or painful.
This figure shows the different options that could happen in screening.

1. **Woman aged 50-70 invited for screening**
2. **Mammography (x-ray of breasts)**
   - Negative Result: **NO CANCER**
   - Positive Result: **CANCER**
   - Positive Result: **Cancerous (fast-growing)**
   - Positive Result: **Cancerous (slow-growing)**
3. **Biopsy required?**
   - **NO CANCER**
   - **CANCER**
4. **Treatment offered?**
   - **YES**
   - **NO**
5. **Outcome**
   - **Woman finds out she does not have cancer**
   - **Woman has a biopsy but no cancer is found**
   - **Woman has a fast-growing cancer and is cured by screening**
   - **Woman receives treatment for cancer to stop it from spreading**

2 weeks after mammogram to results.
Here is a short video explaining what happens in a mammography. This video is available on NHS England’s website should you require any more information or wish to watch it again later.
In the UK, women over the age of 50 are offered screening up until the age of 70. Although the service is free, women who participate in screening incur some cost. For example, they may have to take some time off work, travel to the centre for breast screening, pay for parking and transport, or maybe pay for childcare.

There are **benefits** to breast screening for cancer which means tumours are picked up earlier and the long-term health outcomes of women are improved. However, there are **risks** because it is inconvenient and can mean women undergo procedures unnecessarily.

In this survey we are interested in understanding how women balance these benefits and risks. To do that, we would now like you to complete this survey where you will be shown different screening programmes. You will be asked to choose your preferred programme from a set of possible screening programmes.

**There are no wrong or right answers: we are just interested in your views**.
Risk is a term used to explain the chance that something bad might happen. In breast cancer screening, there is a chance that women might have an unnecessary follow-up after their mammogram. This will not happen to all women, but there is a risk it could happen to some women.

The women shaped figures represent the total number of women having screening. The purple coloured figure shows the one woman who went for screening and had an unnecessary follow-up. This means 1 out of every 100 women screened was followed-up when she didn’t need to be, but 99 women were treated appropriately.

OR

Risk is a term used to explain the chance that something bad might happen. In breast cancer screening, there is a chance that women might have unnecessary follow-up after their mammogram. This will not happen to all women, but there is a risk it could happen to some women.

1%

A risk of 1% means that 1 out of every 100 women screened was followed-up when she didn’t need to be, but 99 women were treated appropriately.
This grid shows that 5 out of every 100 women screened was followed-up when she didn’t need to be, but 95 women were treated appropriately.

OR

Similarly, a risk of 5% means that 5 out of every 100 women screened was followed-up when she didn’t need to be, but 95 women were treated appropriately.
You are going to be asked to choose between screening programmes which differ in terms of:

**Women who will have cancers detected by screening**
This describes the number of women who will have a cancer identified by a mammogram (x-ray of the breast). This describes the number of women, who go for screening over a period of 20 years, that are identified as having a cancer.

**Women who will have unnecessary follow-up**
This shows the number of women requiring a procedure when the lump was not cancerous or would not have caused any harmful symptoms in their lifetime. The unnecessary follow-up includes biopsies and possibly further surgery and chemotherapy.

**Out-of-pocket cost**
This describes the cost to you of participating in a screening programme. There are no plans to introduce charges for screening with the NHS. This cost reflects your out-of-pocket expenses incurred when participating in the screening programme. For example, transport costs, car parking, child care, time off work. In the UK, women will be screened 5 times over a 20 year period (a lifetime programme). We’ll show the cost of the programmes over your lifetime and what that works amounts to per screen.

Each screening programme we show you will be described in terms of these three characteristics.

Each characteristic, however, has four different levels that will change.
Women who will have cancers detected by screening

3% means that if 100 women are screened for 20 years, then 3 women will be identified as having cancer.

7% means that if 100 women are screened for 20 years, then 7 women will be identified as having cancer.

10% means that if 100 women are screened for 20 years, then 10 women will be identified as having cancer.

14% means that if 100 women are screened for 20 years, then 14 women will be identified as having cancer.
Women who will have an unnecessary follow-up

0% means that if 100 women are screened for 20 years, then 0 women will be followed-up unnecessarily.

1% means that if 100 women are screened for 20 years, then 1 woman will be followed-up unnecessarily.

5% means that if 100 women are screened for 20 years, then 5 women will be followed-up unnecessarily.

10% means that if 100 women are screened for 20 years, then 10 women will be followed-up unnecessarily.

20% means that if 100 women are screened for 20 years, then 20 women will be followed-up unnecessarily.

OR

Women who will have an unnecessary follow-up

0% means that if 100 women are screened for 20 years, then 0 women will be followed-up unnecessarily.

1% means that if 100 women are screened for 20 years, then 1 woman will be followed-up unnecessarily.

5% means that if 100 women are screened for 20 years, then 5 women will be followed-up unnecessarily.

10% means that if 100 women are screened for 20 years, then 10 women will be followed-up unnecessarily.

20% means that if 100 women are screened for 20 years, then 20 women will be followed-up unnecessarily.
Out-of-pocket cost to you of screening programme per year

£20 (or £100 over your lifetime)

£50 (or £250 over your lifetime)

£150 (or £750 over your lifetime)

£200 (or £1,000 over your lifetime)
Below is an example of one possible breast screening programme.

<table>
<thead>
<tr>
<th>Programme</th>
<th>3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers detected by screening</td>
<td><img src="image" alt="Graph showing 3%" /></td>
</tr>
<tr>
<td>Women who will have an unnecessary follow-up</td>
<td><img src="image" alt="Graph showing 0%" /></td>
</tr>
<tr>
<td>Out-of-pocket cost to you of screening programme per screen</td>
<td>£20 per screen (€100 over your lifetime)</td>
</tr>
</tbody>
</table>

OR

Below is an example of one possible breast screening programme.

<table>
<thead>
<tr>
<th>Programme</th>
<th>3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers detected by screening</td>
<td><img src="image" alt="Graph showing 3%" /></td>
</tr>
<tr>
<td>Women who will have an unnecessary follow-up</td>
<td><img src="image" alt="Graph showing 0%" /></td>
</tr>
<tr>
<td>Out-of-pocket cost to you of screening programme per screen</td>
<td>£20 per screen (€100 over your lifetime)</td>
</tr>
</tbody>
</table>
Below is an example of the type of question we will ask you.

If these were your only options, which would you choose?

<table>
<thead>
<tr>
<th>Programme A</th>
<th>Programme B</th>
<th>No Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers detected by screening</td>
<td>3%</td>
<td>14%</td>
</tr>
<tr>
<td>Women who will have an unnecessary follow-up</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Out-of-pocket cost to you of screening programme per screen</td>
<td>£20 per screen (£100 over your lifetime)</td>
<td>£50 per screen (£250 over your lifetime)</td>
</tr>
</tbody>
</table>

You simply select the programme which you prefer.

OR

Below is an example of the type of question we will ask you.

If these were your only options, which would you choose?

<table>
<thead>
<tr>
<th>Programme A</th>
<th>Programme B</th>
<th>No Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women who will have cancers detected by screening</td>
<td>3%</td>
<td>14%</td>
</tr>
<tr>
<td>Women who will have an unnecessary follow-up</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Out-of-pocket cost to you of screening programme per screen</td>
<td>£20 per screen (£100 over your lifetime)</td>
<td>£50 per screen (£250 over your lifetime)</td>
</tr>
</tbody>
</table>

You simply select the programme which you prefer.
Please press the "next" arrow below to begin the survey.
If these were your only options, which would you choose? (1 of 1 choice questions)

<table>
<thead>
<tr>
<th>Programme A</th>
<th>Programme B</th>
<th>No Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Women who will have cancers detected by screening</strong></td>
<td>3%</td>
<td>14%</td>
</tr>
<tr>
<td><strong>Women who will have an unnecessary follow-up</strong></td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Out-of-pocket cost to you of screening programme per screen</strong></td>
<td>£20 per screen (£100 over your lifetime)</td>
<td>£50 per screen (£250 over your lifetime)</td>
</tr>
</tbody>
</table>

OR

If these were your only options, which would you choose? (1 of 1 choice questions)

<table>
<thead>
<tr>
<th>Programme A</th>
<th>Programme B</th>
<th>No Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Women who will have cancers detected by screening</strong></td>
<td>3%</td>
<td>14%</td>
</tr>
<tr>
<td><strong>Women who will have an unnecessary follow-up</strong></td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Out-of-pocket cost to you of screening programme per screen</strong></td>
<td>£20 per screen (£100 over your lifetime)</td>
<td>£50 per screen (£250 over your lifetime)</td>
</tr>
</tbody>
</table>
Thank you for providing your choices. You will now be asked some questions about the survey and a few questions about yourself.
How confident are you that you would make the same choices if faced with the situations in real-life?

- Very confident I would make the same choices
- Quite confident I would make the same choices
- Not confident I would make the same choices
- Confident I would make different choices

Please explain your answer
On a scale of 1-5, how easy or difficult did you find making choices between the alternatives?

1 (Very easy)   2 (Neither difficult nor easy)   3 (Neither difficult nor easy)   4 (Very difficult)

Feedback: [Feedback1] [Feedback2] [Feedback3] [Feedback4] [Feedback5]

Please explain your answer
On a scale of 1-5, how easy or difficult did you find the survey to understand?

<table>
<thead>
<tr>
<th>1 (Very easy)</th>
<th>2 (Neither difficult nor easy)</th>
<th>3</th>
<th>4</th>
<th>5 (Very difficult)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback1</td>
<td>Feedback2</td>
<td>Feedback3</td>
<td>Feedback4</td>
<td>Feedback5</td>
</tr>
</tbody>
</table>

Was there anything in the survey that was particularly unclear?

- Yes, I understood everything
- No, there were some things that weren't clear or well explained

Please comment in the box below:

Is there anything else about the survey we could change to make it easier to understand (e.g., phrasing, explanations)?
Did you find yourself making choices based on one or two characteristics of the screening programme?

Yes, I based choices on one or two characteristics
No, I considered ALL characteristics

Would you say you looked at the characteristic women who have a cancer detected?
Please select

Would you say you looked at the characteristic risk of unnecessary treatment?
Please select

Would you say you looked at the characteristic cost?
Please select

If you based your choices on one or two characteristics, why was this? (Please tick all that apply):

- There was just too many characteristics on which to make a decision
- The other characteristic(s) were too complicated or it was unclear what it meant
- Other reason (please explain)
What is your occupational status?

1. Employed full-time
2. Employed part-time
3. Self-employed
4. Unemployed
5. Retired
6. Looking after a home/family
7. Student
8. Freelance or temping
9. Long-term sickness
10. Temporarily laid off

What is your occupation?
(Or previous occupation if retired?)
What is the highest level of education you have obtained?

- No formal qualifications
- 1-4 O-levels/GCSEs
- 5+ O-levels/GCSEs
- NVQs
- A-levels/AS-levels
- Undergraduate degree
- Master's degree
- PhD
- Other formal qualification
What is your ethnic group?
- White British/Irish
- White other
- Mixed
- Black/Black British
- Asian/Asian British
- Other (please specify)
What is your religious group?

- [ ] No religion
- [ ] Christian
- [ ] Buddhist
- [ ] Jewish
- [ ] Hindu
- [ ] Muslim
- [ ] Sikh
- [ ] Other (please specify)
Do you have any children?

- [ ] Background9=1 Yes
- [ ] Background9=2 No
Do you have any daughters?

- Yes
- No
Have you ever had breast screening (a mammogram)?

- ExperienceA=1: Yes
- ExperienceA=2: No
- ExperienceA=3: Don’t know

Have you received follow-up (such as an ultrasound, a repeat mammogram or a biopsy) for breast cancer?

- ExperienceB=1: Yes
- ExperienceB=2: No
- ExperienceB=3: Don’t know
Thinking about the last time you went for breast screening, what was your **main** form of transport to the screening?

Please select as appropriate.

- walked
- cycled
- travelled by bus
- travelled by train/metro
- travelled by private car
- travelled by taxi
- travelled by motorbike
- other transport

Still thinking about your last breast screen, how much did it cost you for your return journey (even if your start and end points were different)?

(If you did not pay any transport fare then please enter 0)

Did you incur any other "costs"?

- I took time off work
- I had to pay for care (for children, relative, friend or other dependent)
- I lost income from my own business
- Other
Have you, or has anyone in your family (blood relative or otherwise) had breast cancer?

- Yes
- No
- Don't know

Do you know of anybody else (e.g. friends or colleagues) who has had breast cancer?

- Yes
- No
- Don't know
How would you describe your current risk of breast cancer?

- Risk 0-1: High or very high
- Risk 0-2: Quite high/above average
- Risk 0-3: Average
- Risk 0-4: Slight or low
- Risk 0-5: No idea

How would you describe your future risk of breast cancer?

- Risk 0-1: High or very high
- Risk 0-2: Quite high/above average
- Risk 0-3: Average
- Risk 0-4: Slight or low
- Risk 0-5: No idea

To what extent are you concerned about your own risk of breast cancer?

- Risk 0-1: Not at all
- Risk 0-2: A little
- Risk 0-3: Quite a lot
- Risk 0-4: Very much
- Risk 0-5: No idea

Is there anything else you would like to tell us about your risk of breast cancer?

---

94.1%
On a scale of 1-5, how much do you agree with the following statements?

"I feel life is very much out of our hands, and fate determines everything"

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="feedback1.png" alt="Feedback" /></td>
<td><img src="feedback2.png" alt="Feedback" /></td>
<td><img src="feedback3.png" alt="Feedback" /></td>
<td><img src="feedback4.png" alt="Feedback" /></td>
<td><img src="feedback5.png" alt="Feedback" /></td>
</tr>
</tbody>
</table>

"I get a thrill when I gamble, even if it's just the lottery"

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="feedback1.png" alt="Feedback" /></td>
<td><img src="feedback2.png" alt="Feedback" /></td>
<td><img src="feedback3.png" alt="Feedback" /></td>
<td><img src="feedback4.png" alt="Feedback" /></td>
<td><img src="feedback5.png" alt="Feedback" /></td>
</tr>
</tbody>
</table>

"Even if something is unlikely, I feel it will probably happen to me"

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="feedback1.png" alt="Feedback" /></td>
<td><img src="feedback2.png" alt="Feedback" /></td>
<td><img src="feedback3.png" alt="Feedback" /></td>
<td><img src="feedback4.png" alt="Feedback" /></td>
<td><img src="feedback5.png" alt="Feedback" /></td>
</tr>
</tbody>
</table>
This next section will ask you some questions to understand how familiar you are with probabilities.

If you are unsure of the answer, you can skip the question.

Imagine I flip a coin 1,000 times. What is your best guess about how many times the coin would fall heads up in 1,000 flips? [__________]

If there was a lottery where the chance of winning £10 is 1%, if 1,000 people bought a ticket, how many people would you expect to win? [__________]

In another lottery, the chance of winning a car is 1 in 1,000. What percentage of tickets in the lottery will win a car? [__________]
You missed some of the previous questions about statistics and probability. Why was this?

☐ Just too hard

☐ We were going to take me too long to answer

☐ Can't see the relevance to the study

☐ Other

0% 100%
Can you think of anything else which might help us understand how you made your choices in the choice task?
Thank you for completing this survey.

If you have been affected by any of the issues raised in this survey, the following organisations may be of assistance:

- National Hereditary Breast Cancer Helpline on 01629 813000
- Breast Cancer Care on 0808 800 6000

If you have any questions or concerns please contact Caroline Veeva on 077 1675 1262 or caroline.veeva@postgrad.manchester.ac.uk.

Manchester Centre for Health Economics, Jean McFarlane Building, University of Manchester, Oxford Road, Manchester M13 9PL
Appendix 3.17: Quantitative results of the DCE from the internet panel in the second pilot study

Table A1: Response statistics

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Icon arrays</th>
<th>Percentages only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed to termination</td>
<td>58</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>Completed choice questions</td>
<td>59</td>
<td>28</td>
<td>31</td>
</tr>
</tbody>
</table>

Table A2: Effect of attributes on choice probabilities and testing for significance of the scale parameter (heteroskedastic conditional logit)

<table>
<thead>
<tr>
<th>Effect of attributes on choice probabilities: pooled model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Detect</td>
<td>0.107*** (0.13)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.013*** (0.01)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.070 (0.00)</td>
</tr>
<tr>
<td>IAP.detect</td>
<td>0.016*** (0.01)</td>
</tr>
<tr>
<td>IAP.risk</td>
<td>-0.028*** (0.01)</td>
</tr>
<tr>
<td>Observations</td>
<td>1914</td>
</tr>
<tr>
<td>Scale term: IAP</td>
<td>7.574** (2.43)</td>
</tr>
</tbody>
</table>

*p<0.05  ** p<0.01  *** p<0.001; Standard errors in parentheses
Appendix 4.1: ResearchNow® response rate in the large internet panel DCE

<table>
<thead>
<tr>
<th></th>
<th>n*</th>
<th>Unique panellists*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emails Selected:</td>
<td>38,251</td>
<td>33,065</td>
</tr>
<tr>
<td>Emails Delivered:</td>
<td>23,082</td>
<td>23,082</td>
</tr>
<tr>
<td>Survey Loads:</td>
<td>2,205</td>
<td>2,205</td>
</tr>
<tr>
<td>Completes:</td>
<td>1,257</td>
<td>1,257</td>
</tr>
<tr>
<td>Screen Outs:</td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>Quota Full:</td>
<td>511</td>
<td>511</td>
</tr>
<tr>
<td>Closed Survey:</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Drop Outs:</td>
<td>336</td>
<td>336</td>
</tr>
<tr>
<td>Response Rate:</td>
<td>9.55%</td>
<td>9.55%</td>
</tr>
<tr>
<td>Qualifying completion rate:</td>
<td>57.01%</td>
<td>57.01%</td>
</tr>
<tr>
<td>Hit Rate:</td>
<td>5.45%</td>
<td>5.45%</td>
</tr>
<tr>
<td>Screenout rate:</td>
<td>3.54%</td>
<td>3.54%</td>
</tr>
<tr>
<td>QuotaFull rate:</td>
<td>23.17%</td>
<td>23.17%</td>
</tr>
<tr>
<td>Drop out Rate:</td>
<td>15.24%</td>
<td>15.24%</td>
</tr>
<tr>
<td>Incidence:</td>
<td>94.16%</td>
<td>94.16%</td>
</tr>
</tbody>
</table>

*Cannot be separated from pilot studies described in Chapter Five

Response rate = (survey loads / emails delivered)
Qualified Compl. rate = (completes / survey loads)
Hit rate = (completes / emails delivered)
Drop out rate = (drop out / survey loads)
Screenout rate = (Screenouts / survey loads)
QuotaFull rate = (Quota Full / survey loads)
Incidence = (completes / (completes + screenouts))
Appendix 4.2: Sample characteristics of respondents who completed the DCE in the final study

<table>
<thead>
<tr>
<th>Age group</th>
<th>Overall (both risk communication formats)</th>
<th></th>
<th></th>
<th></th>
<th>PO</th>
<th></th>
<th></th>
<th>IAP</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Percentage</td>
<td>Cumulative</td>
<td>Frequency</td>
<td>Percentage</td>
<td>Cumulative</td>
<td>Frequency</td>
<td>Percentage</td>
<td>Cumulative</td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>104</td>
<td>10.22%</td>
<td>10.22%</td>
<td>44</td>
<td>8.68%</td>
<td>8.68%</td>
<td>60</td>
<td>11.74%</td>
<td>11.74%</td>
<td></td>
</tr>
<tr>
<td>25-34</td>
<td>103</td>
<td>10.12%</td>
<td>20.33%</td>
<td>51</td>
<td>10.06%</td>
<td>18.74%</td>
<td>52</td>
<td>10.18%</td>
<td>21.92%</td>
<td></td>
</tr>
<tr>
<td>35-44</td>
<td>253</td>
<td>24.85%</td>
<td>45.19%</td>
<td>127</td>
<td>25.05%</td>
<td>43.79%</td>
<td>126</td>
<td>24.66%</td>
<td>46.58%</td>
<td></td>
</tr>
<tr>
<td>45-49</td>
<td>253</td>
<td>24.85%</td>
<td>70.04%</td>
<td>131</td>
<td>25.84%</td>
<td>69.63%</td>
<td>122</td>
<td>23.87%</td>
<td>70.45%</td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td>305</td>
<td>29.96%</td>
<td>100.00%</td>
<td>154</td>
<td>30.37%</td>
<td>100.00%</td>
<td>151</td>
<td>29.55%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>872</td>
<td>86.42%</td>
<td>86.42%</td>
<td>435</td>
<td>86.65%</td>
<td>86.65%</td>
<td>437</td>
<td>86.19%</td>
<td>86.19%</td>
<td></td>
</tr>
<tr>
<td>White British/Irish</td>
<td>872</td>
<td>86.42%</td>
<td>86.42%</td>
<td>435</td>
<td>86.65%</td>
<td>86.65%</td>
<td>437</td>
<td>86.19%</td>
<td>86.19%</td>
<td></td>
</tr>
<tr>
<td>White other</td>
<td>53</td>
<td>5.25%</td>
<td>91.67%</td>
<td>27</td>
<td>5.38%</td>
<td>92.03%</td>
<td>26</td>
<td>5.13%</td>
<td>91.32%</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>14</td>
<td>1.39%</td>
<td>93.06%</td>
<td>7</td>
<td>1.39%</td>
<td>94.82%</td>
<td>9</td>
<td>1.78%</td>
<td>94.48%</td>
<td></td>
</tr>
<tr>
<td>Black/Black British</td>
<td>16</td>
<td>1.59%</td>
<td>94.65%</td>
<td>7</td>
<td>1.39%</td>
<td>94.82%</td>
<td>9</td>
<td>1.78%</td>
<td>94.48%</td>
<td></td>
</tr>
<tr>
<td>Asian/Asian British</td>
<td>45</td>
<td>4.46%</td>
<td>99.11%</td>
<td>22</td>
<td>4.38%</td>
<td>99.20%</td>
<td>23</td>
<td>4.54%</td>
<td>99.01%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>0.89%</td>
<td>100.00%</td>
<td>4</td>
<td>0.80%</td>
<td>100.00%</td>
<td>5</td>
<td>0.99%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Religious Group</td>
<td>382</td>
<td>37.90%</td>
<td>37.90%</td>
<td>188</td>
<td>37.45%</td>
<td>37.45%</td>
<td>194</td>
<td>38.34%</td>
<td>38.34%</td>
<td></td>
</tr>
<tr>
<td>No religion</td>
<td>382</td>
<td>37.90%</td>
<td>37.90%</td>
<td>188</td>
<td>37.45%</td>
<td>37.45%</td>
<td>194</td>
<td>38.34%</td>
<td>38.34%</td>
<td></td>
</tr>
<tr>
<td>Christian</td>
<td>556</td>
<td>55.16%</td>
<td>93.06%</td>
<td>272</td>
<td>54.18%</td>
<td>91.63%</td>
<td>284</td>
<td>56.13%</td>
<td>94.47%</td>
<td></td>
</tr>
<tr>
<td>Buddhist</td>
<td>7</td>
<td>0.69%</td>
<td>93.75%</td>
<td>5</td>
<td>1.00%</td>
<td>94.73%</td>
<td>6</td>
<td>1.20%</td>
<td>94.93%</td>
<td></td>
</tr>
<tr>
<td>Jewish</td>
<td>10</td>
<td>0.99%</td>
<td>94.73%</td>
<td>6</td>
<td>1.20%</td>
<td>94.82%</td>
<td>4</td>
<td>0.79%</td>
<td>95.65%</td>
<td></td>
</tr>
<tr>
<td>Hindu</td>
<td>10</td>
<td>0.99%</td>
<td>95.73%</td>
<td>5</td>
<td>1.00%</td>
<td>95.82%</td>
<td>5</td>
<td>0.99%</td>
<td>96.64%</td>
<td></td>
</tr>
<tr>
<td>Muslim</td>
<td>19</td>
<td>1.88%</td>
<td>97.62%</td>
<td>10</td>
<td>1.99%</td>
<td>98.61%</td>
<td>9</td>
<td>1.78%</td>
<td>98.42%</td>
<td></td>
</tr>
<tr>
<td>Sikh</td>
<td>4</td>
<td>0.40%</td>
<td>98.02%</td>
<td>3</td>
<td>0.60%</td>
<td>97.41%</td>
<td>1</td>
<td>0.20%</td>
<td>98.62%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>20</td>
<td>1.98%</td>
<td>100.00%</td>
<td>13</td>
<td>2.59%</td>
<td>100.00%</td>
<td>7</td>
<td>1.38%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Occupational Status</td>
<td>343</td>
<td>33.99%</td>
<td>33.99%</td>
<td>167</td>
<td>33.27%</td>
<td>33.27%</td>
<td>176</td>
<td>34.71%</td>
<td>34.71%</td>
<td></td>
</tr>
<tr>
<td>Employed full-time</td>
<td>343</td>
<td>33.99%</td>
<td>33.99%</td>
<td>167</td>
<td>33.27%</td>
<td>33.27%</td>
<td>176</td>
<td>34.71%</td>
<td>34.71%</td>
<td></td>
</tr>
<tr>
<td>Employed part-time</td>
<td>211</td>
<td>20.91%</td>
<td>54.91%</td>
<td>114</td>
<td>22.71%</td>
<td>55.98%</td>
<td>97</td>
<td>19.13%</td>
<td>53.85%</td>
<td></td>
</tr>
<tr>
<td>Self-employed</td>
<td>58</td>
<td>5.75%</td>
<td>60.65%</td>
<td>31</td>
<td>6.18%</td>
<td>62.15%</td>
<td>27</td>
<td>5.33%</td>
<td>59.17%</td>
<td></td>
</tr>
<tr>
<td>Unemployed</td>
<td>30</td>
<td>2.97%</td>
<td>63.63%</td>
<td>17</td>
<td>3.39%</td>
<td>65.54%</td>
<td>13</td>
<td>2.56%</td>
<td>61.74%</td>
<td></td>
</tr>
<tr>
<td>Retired</td>
<td>133</td>
<td>13.18%</td>
<td>76.81%</td>
<td>67</td>
<td>13.35%</td>
<td>78.88%</td>
<td>66</td>
<td>13.02%</td>
<td>74.75%</td>
<td></td>
</tr>
<tr>
<td>Looking after home/ family</td>
<td>137</td>
<td>13.58%</td>
<td>90.39%</td>
<td>67</td>
<td>13.35%</td>
<td>92.23%</td>
<td>70</td>
<td>13.81%</td>
<td>88.56%</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>58</td>
<td>5.75%</td>
<td>96.13%</td>
<td>25</td>
<td>4.98%</td>
<td>97.21%</td>
<td>33</td>
<td>6.51%</td>
<td>95.07%</td>
<td></td>
</tr>
<tr>
<td>Freelance/ temping</td>
<td>1</td>
<td>0.10%</td>
<td>96.23%</td>
<td>1</td>
<td>0.20%</td>
<td>97.41%</td>
<td>0</td>
<td>0.00%</td>
<td>95.07%</td>
<td></td>
</tr>
<tr>
<td>Long-term sickness</td>
<td>37</td>
<td>3.67%</td>
<td>99.90%</td>
<td>12</td>
<td>2.39%</td>
<td>99.80%</td>
<td>25</td>
<td>4.93%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Temporarily laid off</td>
<td>1</td>
<td>0.10%</td>
<td>100.00%</td>
<td>1</td>
<td>0.20%</td>
<td>100.00%</td>
<td>0</td>
<td>0.00%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Overall (both risk communication formats)</td>
<td>PO</td>
<td>IAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
<td>-----------------------------------------</td>
<td>------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Have children</strong></td>
<td>Frequency</td>
<td>Percentage</td>
<td>Cumulative</td>
<td>Frequency</td>
<td>Percentage</td>
<td>Cumulative</td>
<td>Frequency</td>
<td>Percentage</td>
<td>Cumulative</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>380</td>
<td>37.33%</td>
<td>37.33%</td>
<td>194</td>
<td>38.26%</td>
<td>38.26%</td>
<td>186</td>
<td>36.40%</td>
<td>36.40%</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>638</td>
<td>62.67%</td>
<td>100.00%</td>
<td>313</td>
<td>61.74%</td>
<td>100.00%</td>
<td>325</td>
<td>63.60%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No formal qualifications</td>
<td>40</td>
<td>3.96%</td>
<td>3.96%</td>
<td>21</td>
<td>4.18%</td>
<td>4.18%</td>
<td>19</td>
<td>3.75%</td>
<td>3.75%</td>
<td></td>
</tr>
<tr>
<td>1-4 O-levels</td>
<td>110</td>
<td>10.90%</td>
<td>14.87%</td>
<td>53</td>
<td>10.56%</td>
<td>14.74%</td>
<td>57</td>
<td>11.24%</td>
<td>14.99%</td>
<td></td>
</tr>
<tr>
<td>5+ O-levels</td>
<td>112</td>
<td>11.10%</td>
<td>25.97%</td>
<td>54</td>
<td>10.76%</td>
<td>25.50%</td>
<td>58</td>
<td>11.44%</td>
<td>26.43%</td>
<td></td>
</tr>
<tr>
<td>NVQs</td>
<td>105</td>
<td>10.41%</td>
<td>36.37%</td>
<td>57</td>
<td>11.35%</td>
<td>36.85%</td>
<td>48</td>
<td>9.47%</td>
<td>35.90%</td>
<td></td>
</tr>
<tr>
<td>A-levels/AS</td>
<td>203</td>
<td>20.12%</td>
<td>56.49%</td>
<td>108</td>
<td>21.51%</td>
<td>58.37%</td>
<td>95</td>
<td>18.74%</td>
<td>54.64%</td>
<td></td>
</tr>
<tr>
<td>Undergraduate degree</td>
<td>270</td>
<td>26.76%</td>
<td>83.25%</td>
<td>129</td>
<td>25.70%</td>
<td>84.06%</td>
<td>141</td>
<td>27.81%</td>
<td>82.45%</td>
<td></td>
</tr>
<tr>
<td>Master’s</td>
<td>88</td>
<td>8.72%</td>
<td>91.97%</td>
<td>48</td>
<td>9.56%</td>
<td>93.63%</td>
<td>40</td>
<td>7.89%</td>
<td>90.34%</td>
<td></td>
</tr>
<tr>
<td>PhD</td>
<td>11</td>
<td>1.09%</td>
<td>93.06%</td>
<td>4</td>
<td>0.80%</td>
<td>94.42%</td>
<td>7</td>
<td>1.38%</td>
<td>91.72%</td>
<td></td>
</tr>
<tr>
<td>Other qualification</td>
<td>70</td>
<td>6.94%</td>
<td>100.00%</td>
<td>28</td>
<td>5.58%</td>
<td>100.00%</td>
<td>42</td>
<td>8.28%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td><strong>Experience mammogram</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>436</td>
<td>43.25%</td>
<td>43.25%</td>
<td>217</td>
<td>43.23%</td>
<td>43.23%</td>
<td>219</td>
<td>43.28%</td>
<td>43.28%</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>569</td>
<td>56.45%</td>
<td>99.70%</td>
<td>284</td>
<td>56.57%</td>
<td>99.80%</td>
<td>285</td>
<td>56.32%</td>
<td>99.60%</td>
<td></td>
</tr>
<tr>
<td>Don’t know</td>
<td>3</td>
<td>0.30%</td>
<td>100.00%</td>
<td>1</td>
<td>0.20%</td>
<td>100.00%</td>
<td>2</td>
<td>0.40%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td><strong>Experienced follow-up</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>148</td>
<td>14.68%</td>
<td>14.68%</td>
<td>75</td>
<td>14.94%</td>
<td>14.94%</td>
<td>73</td>
<td>14.43%</td>
<td>14.43%</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>853</td>
<td>84.62%</td>
<td>99.31%</td>
<td>423</td>
<td>84.26%</td>
<td>99.20%</td>
<td>430</td>
<td>84.98%</td>
<td>99.41%</td>
<td></td>
</tr>
<tr>
<td>Don’t know</td>
<td>7</td>
<td>0.69%</td>
<td>100.00%</td>
<td>4</td>
<td>0.80%</td>
<td>100.00%</td>
<td>3</td>
<td>0.59%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td><strong>Experience of breast cancer in family</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>358</td>
<td>35.52%</td>
<td>35.52%</td>
<td>172</td>
<td>34.26%</td>
<td>34.26%</td>
<td>186</td>
<td>36.76%</td>
<td>36.76%</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>618</td>
<td>61.31%</td>
<td>96.83%</td>
<td>318</td>
<td>63.35%</td>
<td>97.61%</td>
<td>300</td>
<td>59.29%</td>
<td>96.05%</td>
<td></td>
</tr>
<tr>
<td>Don’t know</td>
<td>32</td>
<td>3.17%</td>
<td>100.00%</td>
<td>12</td>
<td>2.39%</td>
<td>100.00%</td>
<td>20</td>
<td>3.95%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td><strong>Experience of breast cancer with friends</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>701</td>
<td>69.54%</td>
<td>69.54%</td>
<td>356</td>
<td>70.92%</td>
<td>70.92%</td>
<td>345</td>
<td>68.18%</td>
<td>68.18%</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>291</td>
<td>28.87%</td>
<td>98.41%</td>
<td>141</td>
<td>28.09%</td>
<td>99.00%</td>
<td>150</td>
<td>29.64%</td>
<td>97.83%</td>
<td></td>
</tr>
<tr>
<td>Don’t know</td>
<td>16</td>
<td>1.59%</td>
<td>100.00%</td>
<td>5</td>
<td>1.00%</td>
<td>100.00%</td>
<td>11</td>
<td>2.17%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td><strong>Perceived current risk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High or very high</td>
<td>17</td>
<td>1.69%</td>
<td>1.69%</td>
<td>6</td>
<td>1.20%</td>
<td>1.20%</td>
<td>11</td>
<td>2.17%</td>
<td>2.17%</td>
<td></td>
</tr>
<tr>
<td>Quite high/above average</td>
<td>86</td>
<td>8.54%</td>
<td>10.23%</td>
<td>41</td>
<td>8.18%</td>
<td>9.38%</td>
<td>45</td>
<td>8.89%</td>
<td>11.07%</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>474</td>
<td>47.07%</td>
<td>57.30%</td>
<td>246</td>
<td>49.10%</td>
<td>58.48%</td>
<td>228</td>
<td>45.06%</td>
<td>56.13%</td>
<td></td>
</tr>
<tr>
<td>Slight or low</td>
<td>217</td>
<td>21.55%</td>
<td>78.85%</td>
<td>102</td>
<td>20.36%</td>
<td>78.84%</td>
<td>115</td>
<td>22.73%</td>
<td>78.85%</td>
<td></td>
</tr>
<tr>
<td>No idea</td>
<td>213</td>
<td>21.15%</td>
<td>100.00%</td>
<td>106</td>
<td>21.16%</td>
<td>100.00%</td>
<td>107</td>
<td>21.15%</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Overall (both risk communication formats)</td>
<td>PO</td>
<td>IAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>----</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Perceived future risk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High or very high</td>
<td>23</td>
<td>2.28%</td>
<td>2.28%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quite high/above average</td>
<td>98</td>
<td>9.73%</td>
<td>12.02%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>499</td>
<td>49.55%</td>
<td>61.57%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slight or low</td>
<td>174</td>
<td>17.28%</td>
<td>78.85%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No idea</td>
<td>213</td>
<td>21.15%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Concern about risk of breast cancer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not at all</td>
<td>132</td>
<td>13.11%</td>
<td>13.11%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A little</td>
<td>556</td>
<td>55.21%</td>
<td>68.32%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quite a lot</td>
<td>162</td>
<td>16.09%</td>
<td>84.41%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very much</td>
<td>93</td>
<td>9.24%</td>
<td>93.64%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No idea</td>
<td>64</td>
<td>6.36%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Task difficulty</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very easy</td>
<td>346</td>
<td>34.19%</td>
<td>34.19%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>299</td>
<td>29.55%</td>
<td>63.74%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neither easy nor difficult</td>
<td>260</td>
<td>25.69%</td>
<td>89.43%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult</td>
<td>97</td>
<td>9.58%</td>
<td>99.01%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very difficult</td>
<td>10</td>
<td>0.99%</td>
<td>100.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Appendix 4.3: Results of a simple logit regression to determine if any covariates were a significant predictor of risk communication format

<table>
<thead>
<tr>
<th>Survey risk version</th>
<th>Parameter estimate</th>
<th>Standard errors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Constant</strong></td>
<td>0.291</td>
<td>(0.56)</td>
</tr>
<tr>
<td><strong>18-24 (base category)</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>25-34</td>
<td>-0.494</td>
<td>(0.34)</td>
</tr>
<tr>
<td>35-44</td>
<td>-0.624*</td>
<td>(0.32)</td>
</tr>
<tr>
<td>45-49</td>
<td>-0.760*</td>
<td>(0.33)</td>
</tr>
<tr>
<td>50+</td>
<td>-0.755*</td>
<td>(0.37)</td>
</tr>
<tr>
<td><strong>No formal qualifications (base category)</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1-4 O-levels</td>
<td>0.272</td>
<td>(0.38)</td>
</tr>
<tr>
<td>5+ O-levels</td>
<td>0.178</td>
<td>(0.38)</td>
</tr>
<tr>
<td>NVQs</td>
<td>-0.070</td>
<td>(0.39)</td>
</tr>
<tr>
<td>A-levels/AS</td>
<td>-0.075</td>
<td>(0.37)</td>
</tr>
<tr>
<td>Undergraduate degree</td>
<td>0.204</td>
<td>(0.36)</td>
</tr>
<tr>
<td>Master’s</td>
<td>-0.042</td>
<td>(0.41)</td>
</tr>
<tr>
<td>PhD</td>
<td>0.758</td>
<td>(0.73)</td>
</tr>
<tr>
<td>Other qualification</td>
<td>0.520</td>
<td>(0.42)</td>
</tr>
<tr>
<td><strong>Employed full-time (base category)</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Employed part-time</td>
<td>-0.281</td>
<td>(0.19)</td>
</tr>
<tr>
<td>Self-employed</td>
<td>-0.105</td>
<td>(0.30)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>-0.382</td>
<td>(0.41)</td>
</tr>
<tr>
<td>Retired</td>
<td>-0.115</td>
<td>(0.26)</td>
</tr>
<tr>
<td>Looking after home/ family</td>
<td>-0.113</td>
<td>(0.22)</td>
</tr>
<tr>
<td>Student</td>
<td>-0.045</td>
<td>(0.37)</td>
</tr>
<tr>
<td>Freelance/ temping</td>
<td>0.000</td>
<td>(.)</td>
</tr>
<tr>
<td>Long-term sickness</td>
<td>0.705</td>
<td>(0.38)</td>
</tr>
<tr>
<td>Temporarily laid off</td>
<td>0.000</td>
<td>(.)</td>
</tr>
<tr>
<td><strong>White British/Irish (base category)</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>White other</td>
<td>-0.093</td>
<td>(0.30)</td>
</tr>
<tr>
<td>Mixed</td>
<td>-0.005</td>
<td>(0.57)</td>
</tr>
<tr>
<td>Black/Black British</td>
<td>0.256</td>
<td>(0.52)</td>
</tr>
<tr>
<td>Asian/Asian British</td>
<td>0.343</td>
<td>(0.50)</td>
</tr>
<tr>
<td>Other</td>
<td>0.368</td>
<td>(0.69)</td>
</tr>
<tr>
<td><strong>No religion (base category)</strong></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Christian</td>
<td>0.042</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Buddhist</td>
<td>-1.179</td>
<td>(0.87)</td>
</tr>
<tr>
<td>Jewish</td>
<td>-0.377</td>
<td>(0.67)</td>
</tr>
<tr>
<td>Hindu</td>
<td>-0.495</td>
<td>(0.83)</td>
</tr>
<tr>
<td>Muslim</td>
<td>-0.569</td>
<td>(0.62)</td>
</tr>
<tr>
<td>Sikh</td>
<td>-1.366</td>
<td>(1.27)</td>
</tr>
<tr>
<td>Other</td>
<td>-0.592</td>
<td>(0.50)</td>
</tr>
<tr>
<td><strong>Have children (dummy)</strong></td>
<td>0.301</td>
<td>(0.16)</td>
</tr>
<tr>
<td>Experienced mammogram (dummy)</td>
<td>0.139</td>
<td>(0.20)</td>
</tr>
<tr>
<td>Experienced follow-up (dummy)</td>
<td>-0.084</td>
<td>(0.21)</td>
</tr>
<tr>
<td>Experienced breast cancer (dummy)</td>
<td>-0.065</td>
<td>(0.15)</td>
</tr>
<tr>
<td>Perceived current risk high (dummy)</td>
<td>-0.059</td>
<td>(0.15)</td>
</tr>
<tr>
<td>Perceived future risk high (dummy)</td>
<td>0.071</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Concerned about risk (dummy)</td>
<td>-0.005</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Survey risk version</td>
<td>Parameter estimate</td>
<td>Standard errors</td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
<td>--------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Found task difficult (dummy)</td>
<td>0.035</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Reported ANA to detect attribute (dummy)</td>
<td>-0.191</td>
<td>(0.41)</td>
</tr>
<tr>
<td>Reported ANA to risk attribute (dummy)</td>
<td>0.306</td>
<td>(0.33)</td>
</tr>
<tr>
<td>Reported ANA to cost attribute (dummy)</td>
<td>0.029</td>
<td>(0.19)</td>
</tr>
<tr>
<td>Observations</td>
<td><strong>1005</strong></td>
<td></td>
</tr>
</tbody>
</table>

* p<0.005; ** p<0.01; ***p<0.001
Appendix 4.4: Latent class analysis results from 6-preference-class (no scale-classes) model for the large internet panel DCE

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Pc-1</th>
<th>Pc-2</th>
<th>Pc-3</th>
<th>Pc-4</th>
<th>Pc-5</th>
<th>Pc-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC (none)</td>
<td>-2.252***</td>
<td>-2.083***</td>
<td>3.801***</td>
<td>0.874***</td>
<td>0.352**</td>
<td>1.612***</td>
</tr>
<tr>
<td></td>
<td>(0.580)</td>
<td>(0.107)</td>
<td>(0.387)</td>
<td>(0.137)</td>
<td>(0.139)</td>
<td>(0.134)</td>
</tr>
<tr>
<td>Detect</td>
<td>0.481***</td>
<td>0.122**</td>
<td>0.033**</td>
<td>-0.003</td>
<td>-0.030*</td>
<td>0.022*</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td>(0.015)</td>
<td>(0.011)</td>
<td>(0.013)</td>
<td>(0.027)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.048***</td>
<td>-</td>
<td>-</td>
<td>0.011*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.014)</td>
<td>(0.008)</td>
<td>(0.010)</td>
<td>(0.019)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.088***</td>
<td>-</td>
<td>-</td>
<td>0.039</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.018)</td>
<td>(0.026)</td>
<td>(0.022)</td>
<td>(0.033)</td>
<td>(0.051)</td>
</tr>
<tr>
<td>Preference class proportions</td>
<td>31.0%</td>
<td>24.1%</td>
<td>18.8%</td>
<td>9.7%</td>
<td>8.3%</td>
<td>8.2%</td>
</tr>
</tbody>
</table>
Appendix 5.1: Systematic review of qualitative research alongside healthcare DCEs

Background
Recent systematic reviews of published DCEs have found the designs are becoming increasingly complex with an increase in the number of choice sets presented and an exponential rise in the number containing complex attributes such as time or risk (de Bekker-Grob et al., 2012; Harrison et al., 2014; Ryan & Gerard, 2003b). The increased complexity of DCEs increases the potential for anomalous or inexplicable choices. As a result, a number of studies have explored the implications for the quantitative analyses; such as the inclusion or exclusion of responses which have failed tests for monotonicity or transitivity and the consequences of attribute non-attendance (Lancsar & Louviere, 2006; Lagarde, 2012). The use of tests for consistency in responses has also increased. A review by de Bekker-Grob et al. (2012) found most studies used some sort of internal validity tests to identify transitivity or continuity in choices. In health DCEs, in particular, the choice scenario presented may be about a good or service very unfamiliar to the respondent and the attributes or levels difficult to understand. Any increases in the cognitive burden of the task could potentially result in poor quality data and should be considered carefully (Hall et al., 2004).

While DCEs involve statistical analysis of quantitative choice data, other research areas, particularly in the social sciences, have used qualitative research methods to investigate people’s beliefs and decision-making processes (Berg, 2007). Common data collection methods used include focus-groups, interviews or field-notes, which are usually transcribed and coded (Silverman, 1993). The process of coding transcripts and analysing the qualitative data depends on the theoretical approach chosen by the researchers (Green & Thorogood, 2004).

Although not frequently employed by economists, there is a growing body of research supporting the use of qualitative methods in health economics (Coast, 1999; Coast et al., 2004). In the context of DCEs, qualitative research methods have a number of potential roles. This could include identifying attributes and/or their associated levels, refining the terminology, cognitive piloting of the survey, exploring the use of aids and vignettes, gaining an understanding of respondents’ decision-making processes, or estimating preference heterogeneity before modelling the data (Coast & Horrocks, 2007; de Bekker-Grob et al., 2012; Kløjgaard et al., 2012).

There is evidence that other stated-preference methods have also benefited from the use of
qualitative research in order to gain a deeper-understanding of their results (Baker et al., 2008). For example, a contingent valuation study used qualitative research methods to explain respondents’ WTP valuations and the reasons motivating their responses (Chilton & Hutchinson, 2003). Likewise, a standard gamble study gained additional insight into people’s understanding of the choice set through qualitative research (Baker & Robinson, 2004). In addition, guidelines for conducting DCEs advocate qualitative methods with specific recommendations for the development of attributes and levels (Bridges et al., 2011b; Coast & Horrocks, 2007; Lancsar & Louviere, 2008). However, there has been no explicit investigation of how well these recommendations regarding the use of qualitative methods in the preparation and testing of DCEs have been translated into practice.

Aims and objectives
The aim of this component of the thesis was to systematically identify all published studies in healthcare that reported the use of qualitative methods to inform the design and/or the interpretation of DCEs. The objectives were to: summarise the proportion of DCEs using qualitative methods; assess the context in which the research was applied; identify the methods and techniques used; and, where possible, appraise the quality of the research conducted.

Methods
This study used systematic review methods as advised by the Centre for Reviews and Dissemination (CRD) to identify all healthcare DCEs published in the last decade (since, and including, a previous systematic review) (de Bekker-Grob et al., 2012). Systematic reviews involve the identification of relevant studies, an appraisal of their quality, and a summary of evidence for a specific research questions in an explicit and methodical manner (Khan et al., 2003). These steps make systematic reviews distinctly different to literature reviews or commentaries which may not be standardised or rigorous in their approach.

The systematic review focussed on identifying DCEs rather than other stated preference methods such as CJA, ACA or contingent valuation because, as discussed in Chapter Two, these methods are grounded in different economic theories and are therefore not directly relevant to this review or the overall thesis (Louviere et al., 2010). In line with previous systematic reviews, this review defined qualitative research methods as any exploration of peoples’ thoughts or feelings through the collection of verbal or textual data (Baker et al., 2008). This definition did not include restrictions on free-text comments in DCE studies (although there is debate about whether these constitute qualitative research methods (O’Cathain & Thomas, 2004)).
An electronic search of Medline (Ovid, 1966 to date) was conducted in June 2012. Although other databases could have been searched, the strategy exactly replicated that of published reviews of DCEs (Ryan & Gerard, 2003b; de Bekker-Grob et al., 2012). The search terms used were: ‘discrete choice experiment(s)’, ‘discrete choice model(l)ing’, ‘stated preference’, ‘part-worth utilities’, ‘functional measurement’, ‘paired comparisons’, ‘pairwise choices’, ‘conjoint analysis’, ‘conjoint measurement’, ‘conjoint studies’, and ‘conjoint choice experiment(s)’. The term ‘conjoint analysis’ was included to identify studies which had used discrete choices rather than those which required respondents to rate or rank alternatives. No search terms were used to directly identify qualitative studies as this was deemed to be too restrictive.

Table A1 shows a summary of the inclusion and exclusion criteria used in this review. The primary inclusion criteria were healthcare related and discrete choices (where respondents do not rank or rate, and there are no adaptive elements to the experimental design). Other literatures, such as environment, transport or food, were also excluded. Non-English articles and reviews, guidelines or protocols were not included. Following the initial screening, if an article could not be rejected with certainty on the basis of its abstract, the full text of the article was obtained for further evaluation. Retrieved papers were excluded if, despite the abstract, they still did not meet the inclusion criteria. Abstract screening was conducted by an initial reviewer (CV) and duplicated by a second reviewer (KP). The reviewers met to discuss results and papers which were disputed for inclusion were retrieved for further assessment. Papers were reviewed a second time to identify any articles relating to the same piece of research, thus limiting the problem of double counting a single study.

Table A1: Inclusion and exclusion criteria for the review.

<table>
<thead>
<tr>
<th>Study focus</th>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment type</td>
<td>Healthcare interventions; healthcare services; healthcare jobs</td>
<td>Environmental; food; transport</td>
</tr>
<tr>
<td>Publication type</td>
<td>Discrete choices; empirical study</td>
<td>Ranking or rating of choices; BWS; clinical trials; guidelines or reviews; discussion papers</td>
</tr>
<tr>
<td></td>
<td>International and British studies; English language; published between 2002 and 2012</td>
<td>Foreign languages</td>
</tr>
</tbody>
</table>

The main foci of the review were to: 1) identify and quantify the proportion of DCEs using qualitative methods; 2) investigate the stages in the DCE at which qualitative research is employed; 3) understand the methods and techniques currently used; and 4) where possible, evaluate the quality of the reporting of research. The studies were initially
categorised into three categories: 1) those which reported no qualitative research; 2) those which contained basic reporting which indicated some qualitative research may have been used; and 3) those which indicated an extensive qualitative component was conducted in direct relation to the DCE. This categorisation identified studies, in category three, which contained sufficient detail for critical appraisal. The categories are defined in Table A2.

Table A2: Categorisation of studies by level of reporting.

<table>
<thead>
<tr>
<th></th>
<th>Aims</th>
<th>Methods</th>
<th>Analysis</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>Clear statement of the purpose of the qualitative research</td>
<td>Indication of the technique used to collect qualitative data</td>
<td>A description of how the qualitative data was examined/ software used</td>
<td>An explanation of the outcomes of the qualitative research</td>
</tr>
<tr>
<td>None</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Basic</td>
<td>✓*</td>
<td>✓*</td>
<td>✓*</td>
<td>✓*</td>
</tr>
<tr>
<td>Extensive</td>
<td>✓</td>
<td>✓</td>
<td>✓*</td>
<td>✓*</td>
</tr>
</tbody>
</table>

*Basic reporting to include at least one from: aims, methods, analysis, results.

Appraisal of studies

If a paper reported information on the aims, methods, analysis and results, they were deemed to contain ‘extensive qualitative’ research, suitable for a formal critical appraisal. The categorisation of studies was initially conducted by CV and repeated by two other researchers (Martin Eden and Eleanor Heather). Using qualitative research methods alongside DCEs is unusual and would not necessarily lead to any ‘question formulation’ but instead could reasonably be conducted with a specific aim (for example, informing attributes or checking the design). Similarly, data collected for informing attributes or levels would need to be reduced to a few key themes in contrast to the traditional ‘expansionary’ nature of qualitative data analysis.

Using pre-existing guides for appraisal of qualitative research methods in a traditional sense may have meant the extensive studies identified by this review would have been judged incorrectly or unfairly. It is, however, crucial that the qualitative research contained in the studies was formally assessed in a standardised and systematic way. Therefore a bespoke appraisal tool was developed to include broader issues which were not included in the traditional tools as advised by the CRD (CRD, 2008).

The tool was developed from a comparison of different appraisal forms (See: Critical Appraisal Skills Programme (CASP), 2006; Dixon-Woods, 2004; Long & Godfrey, 2004; NCDDR, 2004; Popay, Rogers, & Williams, 1998; Walsh & Downe, 2006) as recommended in the CRD guidelines for systematic reviews of qualitative research (CRD, 2008). In addition to the tools suggested by the CRD, the Joanna Briggs Institute
Qualitative Assessment and Review Instrument (JBI QARI see Pearson & Field (2007)) was also included in the comparison on the recommendation of experienced qualitative researchers in the department (personal communication with Dr Gavin Daker-White in January 2012).

As a starting point, the identified critical appraisal tools were summarised in a table to enable direct comparison of the suggested quality assessment criteria and coverage of the existing generic appraisal tools (the table can be found in Appendix 5.4). The iterative process involved two researchers (CV and Martin Eden) independently reviewing the tabulated list of quality assessment criteria and selecting criteria that matched the requirements for appraising the use of qualitative research methods in stated preference studies. The aim of this process was to identify commonalities and differences in the assessment criteria in a structured format. From this table, the relevance of each tool and quality appraisal criteria in relation to stated preference studies was established through an independent initial assessment and subsequent discussion by the two researchers to achieve consensus. Problems and factors discovered when testing the existing tools were used to focus these discussions and identify the relevance of the appraisal criteria for use in stated preference studies.

**Results**

*Search results*

One hundred and twenty four studies were already identified by a previous systematic review (de Bekker-Grob et al., 2012). The search resulted in 501 titles and abstracts since the previous review (2008 onwards) and 208 full papers were retrieved for further assessment and 148 papers met the inclusion criteria. Figure A1 shows the stages involved in screening and the reasons for rejection of the excluded papers using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) review guide (Moher et al., 2009). Therefore, in total, there were 277 papers included in the final review, which related to 254 empirical studies (because some studies were reported in more than one paper).

Each study was assigned a specific identity (ID). Previous reviews by de Bekker-Grob et al. (2012) and Ryan & Gerard (2003) also labelled studies with an ID which referred to either a single paper or group of papers of the same DCE. The ID also describes the amount of qualitative research reported in the paper(s) with the prefix A, B and C referring to no qualitative research reported, basic qualitative research reported or extensive qualitative research reported, respectively. The identified studies and associated ID are
presented in Appendix 5.6.

Sixty two of the screened papers were rejected by CV and KP at the first stage because they were not related to healthcare. A further 232 papers were rejected because they did not contain an example of an empirical DCE (36 involved rating or ranking the alternatives; 11 were ACA studies; 22 were guidelines or overviews to conducting DCEs; and three related to prospective studies which had not begun). Ten of the papers were not in English, 27 were already covered by the previous systematic review, and 22 were duplicated in the search results. A detailed description of the reasons for rejection can be found in Appendix 5.3.
Figure A1: Flow of studies through the systematic review

- 501 initial references
  - 21 duplicates
  - 27 covered by previous review
  - 178 not empirical study
  - 57 not health care
  - 10 not English

- 208 references
  - 56 not empirical study
  - 4 not health care

- 148 references for detailed review
  - 124 papers from previous reviews
  - 5 papers identified in data extraction

- 277 final papers
  - 254 studies
Critical appraisal tool

The existing appraisal tools covered eight broad themes: 1) background and overview; 2) description of the context; 3) sampling methods; 4) data collection; 5) reflexivity (or a consideration of how the investigators themselves can influence qualitative research); 6) ethical aspects; 7) data analysis; and 8) presentation of results. Within these eight themes, a set of 32 specific criteria were identified. Each tool differed in terms of the number of criteria included, ranging from a relatively small number of seven or eight criteria up to a more extensive list containing 44 items.

These existing tools were deemed inappropriate for a number of reasons. They failed to distinguish between the different study designs and the specific theoretical approaches for mixed methods studies; the appraisal tools were not validated for all qualitative approaches and the specific application for this thesis was not considered by any existent tool. Discrepancies in criteria were evident when the available appraisal tools were considered as a whole. There was also no ‘best’ all-rounder, resulting in a trade-off in some criteria (for example, emphasising the importance of congruity in the JBI QARI tool at the expense of detailed information about the analysis undertook as covered in CASP).

These existing tools failed to take into account the style of published papers in health economics journals. Stated preference studies tend to be published in health economics journals and are therefore not subject to the same rigorous peer-review for the qualitative component. Consequently, the details required about the philosophical foundations as found in the Long tool (Long & Godfrey, 2004), would penalise some good quality qualitative research in a stated preference study. Similarly, some tools (such as Long & Godfrey (2004) and Popay et al. (1998)) placed too much emphasis on technicalities such as the theoretical framework, the epistemological and ontological foundations, and evidence of conceptual adequacy. Details of these criteria may not be sufficiently explained in a stated preference paper where readers are unlikely to be familiar with qualitative philosophies or terms. To avoid penalising for a ‘tick-box’ exercise, these criteria were deemed unimportant.

In some respects the identified tools were deemed too specific; however, there were also some key criteria missing. When testing the tools on example papers it became apparent that additional criteria should be added with two key concepts in mind: 1) qualitative research methods are often employed in the design phase of stated preference studies and 2) qualitative work is conducted in this area in order to understand and provide evidence
for validity of findings from stated preference studies. These two concepts both have implications for the quantitative findings: 1) reducing the data so that a few key attributes and levels are identified and that these are framed and presented in a way that respondents can easily understand; and 2) determining the extent to which results are in accordance with underlying economic theory and rationale for choice-based studies, and how this impacts on model selection, econometric tests or assumptions.

The bespoke tool was developed using an iterative process. As part of this trialling process additional criteria deemed important but absent from existing tools were added. These criteria were based on emphasising the importance of determining whether or not the qualitative component had explicitly been added to the study with the objective of informing study design and/or understanding of the stated preference study was considered paramount. Consequently, criteria were included in the tool which seek to highlight whether a clear intention and/or application of qualitative findings is discernible. Preference elicitation methods require multi-level synthesis of both qualitative and quantitative data which are often collected as separate streams of textual and choice data, respectively. How these inform each other in the development of the survey or the interpretation of results is unique and was not adequately covered in the existing tools. Although CASP (2006) asks: “How valuable is the research?” and the Long & Godfrey (2004) checklist covered “What are the implications for policy?” it was not believed these questions sufficiently covered studies who used the qualitative component to progress the DCE design. For example, how were many themes reduced to just a few attributes or levels?

Therefore the new appraisal tool allowed for parallel, multi-level synthesis of the qualitative data into their study and a clear description of how this was achieved was also considered to be a key criterion. How the results helped the design or interpretation of the stated preference study specifically was identified as an important consideration. A clear explanation of the consequential interpretation and influence was also included as a new criterion in the bespoke tool.

A further key recommendation to emerge from the development discussions was that the appraisal tool should be designed so as to encourage users to fully consider each criterion within a ‘checklist’ format. Therefore, the existing and additional criteria were formed into questions under a set of key headings. A researcher using the new critical appraisal tool could then elicit a yes/no answer for each criterion, which was then supported by an
A preliminary version of the tool was tested in a pre-selected sample of published DCEs studies which were in known existence before the review commenced (Cheraghi-Sohi et al., 2007; Ryan et al., 2009). Data were extracted from the published articles by two researchers (CV and Martin Eden) in accordance with the proposed quality assessment criteria of the appraisal tool. Face-to-face meetings between researchers were used to discuss the feasibility and acceptability of using the appraisal tool and further refinements were applied as a result of discussions. Appendix 5.5 shows the final version of the bespoke appraisal tool, and Appendix 5.9 presents an article describing the tool currently submitted to the journal Health Economics.

**Identified DCEs**

Details about the included studies and tabulation of the extracted data can be found in Appendix 5.7.

There was an exponential increase in the number of DCEs published over time, with over half of the studies (n=154, 56%) published since 2009. Half of the DCEs identified by this review were published in health services research journals such as Health Policy and Planning; Health Expectations; and Social Science and Medicine (n=139, 50%). A third were published in specialised medical journals (n=88) with the most popular areas being obstetrics and gynaecology (n=13, 15%), respiratory diseases (n=13, 15%) and oncology (n=9, 10%). The remaining articles were covered by general medical journals (n=31, 11%) or other areas (such as risk analysis). A full breakdown of journals publishing the DCEs can be found in Appendix 5.7.

Over half the DCEs published were conducted in Europe (n=186, 56%) and a quarter of the DCEs identified were carried out in the UK (n=84, 25%). Other popular countries included the United States of America (USA) (n=49, 15%), the Netherlands (n=38, 11%), Australia (n=26, 8%) and Canada (n=19, 6%). Nineteen studies (8%) asked residents of multiple countries to complete their DCE and two studies gathered data from delegates at an international conference thus incorporating a wide range of nationalities (A7, A73). Of the studies conducted within the UK (both solely and those looking at multiple countries) most (n=43, 51%) had a focus on primary care services or conditions usually treated by GPs. The settings of other countries were not included in this review due to inconsistencies in the definitions of ‘primary care’ across international healthcare systems.
For example, in some European countries, gynaecology is classified as a primary care service as women can attend a community clinic without referral and as a first point of contact.

*Reporting of qualitative research methods*

Overall, 111 (44%) studies did not report the use of any qualitative research methods; 114 (45%) reported minimal data on the use of qualitative methods; and 29 (11%) reported, or explicitly cited, the extensive use of qualitative methods. The trends in publishing qualitative research generally reflected the increases in the overall number of DCEs as demonstrated by Figure A2.

*Figure A2: Trends in DCE publishing over time*

![Graph showing trends in DCE publishing over time](image)

<table>
<thead>
<tr>
<th>Year</th>
<th>Overall*</th>
<th>None</th>
<th>Basic</th>
<th>Extensive</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2002</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2003</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2004</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2005</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2007</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2008</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2009</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2011</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2012</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

+2012 incomplete due to the year of search.

*Overall includes papers rather than studies.

None refers to studies which indicated no qualitative research.

Basic refers to studies which indicated some qualitative research methods were employed.

Extensive refers to studies which reported the qualitative component of their research in detail.

As shown in Figure A3, a variety of applications of qualitative research were identified. Applications included: the selection of attributes and/or levels (n=95, 66%); piloting the DCE (n=26, 18%); and understanding respondents’ decision-making processes (n=4, 3%). Popular qualitative research methods included focus groups (n=64, 45%) and interviews.
(n=108, 76%). Semi-structured interviews (n=25, 17%), structured interviews (n=5, 3%) and cognitive debriefing techniques (n=12, 8%) were the most common approaches.

Figure A3: Potential use of qualitative research methods with DCEs

The next section will discuss in detail how these results were reported with either a basic or extensive description of the work undertaken.
Basic qualitative research

The following section describes the qualitative component of studies which indicated that they had used qualitative research methods alongside their DCE study but provided insufficient details to appraise the study.

Level of reporting

Almost all authors who reported using some qualitative research did so by stating in the methods section of the paper the nature of the qualitative component of their research (n=113, 99%). One study only mentioned that qualitative research was conducted but gave no indication of the collection methods used (B78), and two studies indicated that qualitative research was conducted in the design of the DCE but gave no further details about the context (B24, B34).

Sampling

The 16 studies (14%) which reported using qualitative research methods prior to implementing the DCE survey chose to sample from a different population to that finally used in their DCE study. Most often the qualitative research used a sample of researchers or healthcare professionals before distributing the final DCE to patients or the public.

Context

Almost all (n=113, 99%) of the studies which reported basic qualitative research reported using it before the DCE was implemented, either in the design or piloting phase. None of the studies reported using qualitative research methods alongside the final survey, although three studies reported using qualitative research at the end of the DCE to attain additional information on preferences (B83, B32, B25).

A variety of applications of qualitative research methods were identified. In the design of the DCE, researchers were most commonly seeking to identify attributes and/or assign levels (n=70, 61%), or validate attributes and/or levels identified through other methods (n=31, 27%). Researchers also used qualitative research methods more specifically to check terminology, vignettes and descriptions (n=9, 8%) and to confirm translations (n=2, 2%).

After the design phase, some studies also reported using qualitative research methods in the piloting of the DCE (n=24, 21%). In the pilot stage, the methods were specifically used to check for decision strategies (n=1, 1%) and also to determine an appropriate sample for
the final DCE. For example, study B5 used interviews to determine an appropriate age range for the final DCE. One study (B49) used the qualitative data acquired in the piloting stage to estimate preference heterogeneity and thus predict an appropriate model for the choice data, and another study (B11) used the qualitative research to predict the signs of the coefficients.

Data collection methods
Within the studies that had used basic reporting of qualitative methods, the most popular approach to qualitative data collection was interviews (n=89, 78%), comprising a range of techniques including structured and semi-structured interviews and focus groups. Ten studies also employed cognitive interviews which included debriefing questions at the end of the task as well as a verbal protocol analytical technique called ‘think-aloud’. Focus groups were also another popular approach of data collection (n=50, 44%).

Analysis of qualitative data
Although a crucial step in drawing reliable and valid results from the qualitative data, only a minority of studies mentioned anything about the approach to the analysis of the qualitative data (n=15, 7%). Of these 15 studies, five reported using content analysis (B22, B25, B35, B51, B106) and two studies using framework analysis (B29, B31). Other analytical approaches included the use of grounded theory methods such as the constant comparative method (B33) and open-ended coding (B113). Three studies detailed the use of specialist qualitative software: two used NVivo® (B25, B65); and one used Atlas.ti® (B112). Table A3 shows the methods and context of the qualitative research reported in the DCEs.
Table A3: Summary of methods and context of the qualitative research reported in DCEs

<table>
<thead>
<tr>
<th>Methods</th>
<th>Studies (n)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive interviews</td>
<td>10</td>
<td>9%</td>
</tr>
<tr>
<td>Focus groups</td>
<td>50</td>
<td>44%</td>
</tr>
<tr>
<td>Free-text comments</td>
<td>2</td>
<td>2%</td>
</tr>
<tr>
<td>General interviews</td>
<td>59</td>
<td>52%</td>
</tr>
<tr>
<td>Semi-structured interviews</td>
<td>12</td>
<td>11%</td>
</tr>
<tr>
<td>Structured interviews</td>
<td>5</td>
<td>4%</td>
</tr>
<tr>
<td>Telephone interviews</td>
<td>3</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>141</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Context**

<table>
<thead>
<tr>
<th>Context</th>
<th>Studies (n)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>To add additional information about preferences to support results</td>
<td>4</td>
<td>4%</td>
</tr>
<tr>
<td>To check for decision strategies</td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>To check terminology and descriptions</td>
<td>9</td>
<td>8%</td>
</tr>
<tr>
<td>To check translations</td>
<td>2</td>
<td>2%</td>
</tr>
<tr>
<td>To identify attributes and/or levels</td>
<td>70</td>
<td>61%</td>
</tr>
<tr>
<td>To inform general DCE design</td>
<td>3</td>
<td>3%</td>
</tr>
<tr>
<td>To inform the quantitative analysis</td>
<td>2</td>
<td>2%</td>
</tr>
<tr>
<td>To inform the sampling strategy</td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>To pilot/pre-test the DCE</td>
<td>24</td>
<td>21%</td>
</tr>
<tr>
<td>To validate attributes and/or levels identified through other methods</td>
<td>31</td>
<td>27%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>147</strong></td>
<td></td>
</tr>
</tbody>
</table>

*Some studies used multiple methods or employed qualitative research methods in multiple contexts.

Extensive qualitative research

The following sections describes the qualitative component of studies which provided extensive details about the qualitative component of their study stating the aims, methods, analysis and results either in the text or with a citation to other work.

Level of reporting

Seven studies extensively described the use of qualitative research within the main text of the paper. Twenty-two further studies were identified as having conducted extensive qualitative research by checking the references to the qualitative component of the work. The cited qualitative research was published in a variety of sources including peer-reviewed journals (n=17) and National Institute for Health Research (NIHR) commissioned reports (n=4).

Context

Most studies (n=25, 86%) reported the use of qualitative research methods to identify attributes and/or levels for use in the DCE. Three studies (C19, C9 and C4) used qualitative research methods to understand more about how respondents completed the choice task presented. Two studies (C19 and C12) also used the qualitative methods as a
complement to the quantitative analysis. Other studies used qualitative research methods to pilot and refine the survey (C10, C15).

Data collection methods
The most common data collection approach was interviews (n=18, 62%). These interviews were mostly semi-structured (n=12, 41%) and face-to-face, although two studies used telephone interviews (C7 and C18). Of the three studies using qualitative research to understand more about how people completed the DCE task, two of these (C4 and C9) used an approach called think-aloud.
A number of studies also used focus groups or group interviews (n=12, 41%), and four studies used a combination of focus groups and interviews in their qualitative study (C22, C24, C14 and C27). One study (C10) used the results of an ethnographic direct observational study to identify attributes and levels for the DCE, and used semi-structured interviews to refine the training materials and descriptions.

Analysis of qualitative data
Most studies simply stated in the paper that they used thematic analysis (n=10, 34%) or content analysis (n=5, 17%) to categorise the qualitative data collected. One study also reported the use of a ‘latent’ content approach to discover underlying themes (C25). A type of thematic analysis, thematic synthesis was reported by C29 which involves a more explicit refinement of themes (possibly from multiple studies) and is an approach in line with reducing the qualitative data to develop a few attributes and levels.
Other analytical approaches included framework analysis (n=3, 10%) and a related approach to qualitative analysis called charting (C4). Seven studies used some constant comparative analysis (n=4, 14%) or open-coding (n=3, 10%) at least in the initial stages. Two studies (C18 and C17) used interpretative phenomenological analysis (IPA) which often takes an open-coding approach rather than relying on pre-existing themes or frameworks.

The type of software used was not always reported but the most commonly reported packages were NVivo® (n=4, 14%) and Atlas.ti® (n=2, 7%). One study (C8) also used NUDist®, a software related to NVivo®. The citation of the qualitative research (either the main-text of the DCE or a previous publication); the application; the methods employed; and the analysis conducted are described in Table A4. The results of the qualitative analysis and how these influenced (or did not influence) the DCE study was unique to each paper and cannot readily be quantified or summarise.
<table>
<thead>
<tr>
<th>ID, country, cited research</th>
<th>Methods</th>
<th>Context</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>C16 (UK)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C18 (Multi-country)</td>
<td>Open-ended interviews</td>
<td>To identify attributes and levels</td>
<td>Frequency analysis and IPA</td>
</tr>
<tr>
<td>C19 (Germany)</td>
<td>Unstructured interviews</td>
<td>To understand how respondents complete the choice task (trading behaviour)</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C9 (UK)</td>
<td>Think-aloud interviews</td>
<td>To understand how respondents complete the choice task (trading behaviour)</td>
<td>Coded using a literature derived framework</td>
</tr>
<tr>
<td>C5 (USA)</td>
<td>Semi-structured interviews</td>
<td>To identify attributes and levels</td>
<td>Constant comparative method in Atlas.ti® and ‘open-coding’</td>
</tr>
<tr>
<td>C10 (UK)</td>
<td>Ethnographical observation study</td>
<td>To identify attributes and levels</td>
<td>Constant comparative method in Atlas.ti®</td>
</tr>
<tr>
<td>C17 (Canada)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>IPA content analysis</td>
</tr>
<tr>
<td>C3 (Australia)</td>
<td>Semi-structured interviews</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C6 (Multi-country)</td>
<td>Semi-structured interviews</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C13 (USA)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C20 (USA)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>Content analysis using Atlas.ti®</td>
</tr>
<tr>
<td>C21 (Ghana)</td>
<td>Semi-structured interviews</td>
<td>To identify attributes and levels</td>
<td>Content analysis using NVivo®</td>
</tr>
<tr>
<td>C11 (UK)</td>
<td>Focus groups</td>
<td>To validate attributes and levels</td>
<td>Content analysis</td>
</tr>
<tr>
<td>C1 (USA)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C28 (Australia)</td>
<td>Semi-structured interviews</td>
<td>To identify attributes and levels</td>
<td>Thematic synthesis</td>
</tr>
<tr>
<td>ID, country, cited research</td>
<td>Methods</td>
<td>Context</td>
<td>Analysis</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>C29 (Australia)</td>
<td>Semi-structured interviews</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C22 (UK)</td>
<td>Semi-structured interviews, Focus groups</td>
<td>To identify attributes and levels</td>
<td>Constant comparative analysis</td>
</tr>
<tr>
<td>C24 (UK)</td>
<td>Interviews</td>
<td>To identify attributes and levels</td>
<td>Theme-based code-book in NVivo®</td>
</tr>
<tr>
<td>C4 (UK)</td>
<td>Think-aloud interviews</td>
<td>To understand how respondents complete the choice task (trading behaviour)</td>
<td>Charting approach</td>
</tr>
<tr>
<td>C7 (Australia)</td>
<td>Group interviews</td>
<td>To identify attributes and levels</td>
<td>Latent content analysis</td>
</tr>
<tr>
<td>C8 (UK)</td>
<td>Semi-structured interviews</td>
<td>To identify attributes and levels</td>
<td>Open coding (moving to structured) using NUDist software and NVivo®</td>
</tr>
<tr>
<td>C2 (UK)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C12 (UK)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>Inform quantitative (subgroup) analysis</td>
</tr>
<tr>
<td>C23 (UK)</td>
<td>Interviews</td>
<td>To identify attributes and levels</td>
<td>Thematic analysis</td>
</tr>
<tr>
<td>C25 (UK)</td>
<td>Group interviews</td>
<td>To identify attributes and levels</td>
<td>Latent content analysis</td>
</tr>
<tr>
<td>C15 (Australia)</td>
<td>Interviews</td>
<td>Piloting and pre-testing</td>
<td>Content analysis</td>
</tr>
<tr>
<td>C14 (UK)</td>
<td>Focus groups</td>
<td>To identify attributes and levels</td>
<td>Framework analysis</td>
</tr>
<tr>
<td>C27 (Vietnam)</td>
<td>Semi-structured interviews, Focus groups</td>
<td>To identify attributes and levels</td>
<td>Coding in NVivo®</td>
</tr>
</tbody>
</table>

'See also' directs to the cited qualitative research related to the DCE study.
**Discussion**

This systematic review identified all published health DCEs since 2003 and assessed the qualitative research reported in the paper. A total of 254 relevant studies were identified. A key finding of the systematic review was that a large proportion of DCE studies either do not conduct, or fail to explicitly report, using qualitative research methods. Some studies did acknowledge the lack of qualitative research was a study limitation.

Although the systematic review found that the most common application of qualitative research was to select attributes and levels for use in the DCE, other applications were also identified. Qualitative research was also frequently used in the pre-testing or piloting the DCE survey, and for refining or checking terminology. The review also found some studies using qualitative research methods in other applications. For example, to predict preference heterogeneity, select and specify a regression model, to identify the motives behind ‘irrational responses’, or to specifically test for breaks in the key axioms which support DCEs as a method (Kjaer & Gyrd-Hansen, 2008; Kjaer et al., 2006; Ryan et al., 2009). In light of these broad ranging applications, it is apparent that qualitative research methods are being used in ways beyond the advice of general conduct instructions or even specific guidelines.

The systematic review was the first systematic review to conduct an in-depth evaluation of the different applications, methods and quality of qualitative research reported in DCEs. Previous reviews of health DCEs have provided overviews of the design and analyses or assessed their content (Ryan & Gerard, 2003b; de Bekker-Grob et al., 2012; Harrison et al., 2014). Although de Bekker-Grob et al. (2012) gave a brief summary of qualitative research conducted alongside DCEs, but no details on the aims, methods or analysis were extracted from their papers.

Since the systematic review was conducted, another systematic review of DCEs in health has been published. Clark et al. (2014) also updated the review de Bekker-Grob et al. (2012) and running an identical search strategy, but again there was a more general focus and qualitative research methods were only briefly summarised.

Furthermore, the review by Clark et al. (2014) suffered from a number of serious limitations. The review focussed on the literature published in 2009-2012 but missed several DCE studies published in 2008 after the review by de Bekker-Grob et al. (2012). For example, studies A40, B43, B46 and B47 identified in the systematic review reported...
above were not included. Whilst missing some studies, the review by Clark et al. (2014) also suffered from double-counting as it included papers (such as B58 and A64) covered by the earlier review that they sought to update (de Bekker-Grob et al., 2012). Additionally, Clark et al. (2014) included some conjoint rating studies which did not meet their own inclusion criteria (such as Waltzman et al. 2011; and Bederman et al. 2010). A BWS type-2 study was also included in an apparent error (see Gunther et al. 2010).

Most importantly, the results of the systematic review identified almost sixty DCE studies not identified by Clark et al. (2014). A list of the studies included by this systematic review but missed by Clark et al. (2014) can be found in Appendix 5.8. It is unknown why these papers were not included by Clark et al. (2014).

**Limitations**

A limitation of the systematic review was the focus on papers recorded in one database, Medline. This search strategy was chosen because it updated a previously published review by de Bekker-Grob et al. (2012) and replicated their study. The authors of the review chose Medline as other databases such as Pubmed or Embase identified duplicate papers rather than missing studies. The authors state in their review: “It was expected to identify the large majority of the health-related DCE studies published during the period” (p. 146). A second limitation of the review is the lack of secondary data extraction. Abstract screening was repeated by a second researcher and the categorisation of papers into the extent of use of qualitative research methods was conducted by two other researchers. Final data extraction was conducted by CV only.

One of the most important limitations of the systematic review in this thesis was the reliance on what was reported in the published paper. It could be that rigorous qualitative research was being extensively conducted but the details were never reported in the final journal article, perhaps because of word restrictions. As a rebuttal to this limitation, a follow-up survey to authors of DCEs included in the review was conducted and is presented in Appendix 5.2.

**Conclusion**

The systematic review identified two papers (Cheraghi-Sohi et al., 2007; Ryan et al., 2009) which used concurrent verbal protocol analysis (called think-aloud) to understand more about people’s choices. The think-aloud method was considered appropriate to answer the key research questions about understanding how respondents’ complete a DCE.
There was a paucity of detail about the analysis of the qualitative data collected, however, when described, most studies used some sort of thematic analysis, which was most often a pre-defined framework to code the collected data from which themes were developed using NVivo® software. Therefore the analysis of the think-aloud data could follow a similar approach.

**Bibliography**


Kennedy, A., Gately, C. and Rogers, A. (2005). Examination of the implementation of the Expert Patients Programme within the structures and locality contexts of the NHS in England (PREPP study) - Repoer II. National Primary Care Research and Development Centre.


O’Cathain, A. and Thomas, K.J. (2004). ‘Any other comments?’ Open questions on questionnaires - a bane or a bonus to research? BMC medical research methodology. 4. 25.


Delivery and Organisation Research and Development.


Appendix 5.2: Survey to authors of published healthcare DCEs

Background
The systematic review presented Appendix 5.1 was constrained by its reliance on the details reported in published academic papers. It is well established that a bias exists in both medicine and social science, with studies exhibiting positive or interesting results being more likely to be published (Franco et al., 2014; Easterbrook et al., 1991) A survey was, therefore, designed to elicit information from authors of published DCEs in health about their views on the use of qualitative research methods alongside quantitative analytical methods.

Aims and objectives
The aim of this study was to try and reduce the effects of reporting bias through directly eliciting researchers’ views and experience of conducting qualitative research alongside DCEs in a healthcare setting. The objectives of this study were to reveal more about researchers’ experience of using qualitative research methods; their opinion of the usefulness of qualitative approaches in this context; and any explanation they had for the poor level of reporting found in the systematic review.

Methods
Study sample
Authors who indicated, in a published study, that they had used qualitative research methods but only reported basic details were identified as the most appropriate sample. In choosing authors who were known to have conducted qualitative research would generate interesting results and likely achieve a higher response rate than contacting authors who did not indicate they used qualitative research methods. There seemed little advantage to contacting authors of the studies which reported extensive details of the qualitative component to their research, as a comprehensive account of this aspect of their study had already been captured in the paper and appraised using the bespoke tool.

A total of 114 studies reported basic use of qualitative research methods and all authors were included in this review. As some corresponding authors had multiple studies included in the review, 91 individual authors were sampled.

Survey design
All of the journal articles provided an email address for the corresponding author. Therefore, the most feasible method of contacting authors and eliciting their views was to use an online survey. Whilst other methods such as telephone interviews or a postal survey
could have been used, acquiring other contact details for all authors and given that a large proportion were abroad would have proved difficult. Therefore an online self-administered survey was developed.

The survey was designed using Sawtooth software (Sawtooth, 2012) and hosted on The University of Manchester’s secure server. The final survey consisted of 12 questions and comprised a mix of closed-ended and free-text comment boxes. The questions asked authors about their experience and their opinion of: 1) using qualitative research methods alongside DCEs; and 2) communicating the qualitative work they conducted in a journal article. Additional questions included self-assessment of their and co-authors’ expertise in qualitative research, the number of DCEs they had conducted, and whether they agreed with the key finding of the systematic review that qualitative research is not well reported in healthcare DCEs.

Pilot study
A preliminary study was devised by CV with input from the supervisory team and was piloted with researchers (n=3) experienced with DCEs but not included in this review (because their DCEs were unpublished or in non-health subjects). The pilot study identified that acquiring as much information about the authors’ views was crucial. Therefore the preliminary questionnaire was amended so authors had the opportunity to add further comments to all of their answers, and at the end of the survey they were encouraged to highlight anything that was missing but important in a free-text comment box. The final survey can be found in Appendix 5.10.

Data collection
The email details and the corresponding authors’ details were entered into a spreadsheet and double checked with the public online profiles of authors to ensure accuracy. When an email was returned as ‘undeliverable’, a link was sent through alternative means (such as Academia.edu, Researchgate and LinkedIn profiles). As a last resort when no online profile existed, an alternative author listed on the relevant paper was contacted.

The authors (n=91) were invited to participate via an email (or electronic message) which explained the systematic review and included a brief abstract covering the background, methods and results. The email contained a unique link to identify authors who had failed to either accept or decline the invitation. These authors were sent a reminder after 21 days. For the purpose of the reminder emails, the questionnaire respondent could be identified through the unique link; however, all responses were treated as anonymous. Templates of
the initial email and the reminder email can be found in Appendix 5.11 and 5.12, respectively.

Data analysis
The survey data were downloaded from the online server and analysed in Excel® (Microsoft, 2010). The analysis involved simple production of descriptive statistics for each of the questions and the generation of bar graphs for a visual summarisation. The authors’ free-text comments were not thematically analysed because of the limited textual data available (some authors chose not to comment). Instead these free-text comments were collated into a word document and discussed with the supervisory team to generate core themes.

Results
After the first email sent on 1st May 2013, 38 authors completed the survey (an initial response rate of 42%). A total of 53 authors who had not responded to the initial email were then sent a reminder after three weeks. Four authors declined to take part (for reasons such as one author had not practiced in the field for a few years so could not sufficiently recall their experiences; another was a statistician who had only been involved with the DCE analysis). The questionnaire closed on the 30th June 2013, with a total of 53 completed or partially complete responses, resulting in an overall response rate of 58%. In addition, examples from the free-text comment box are used to provide additional contextual data provided by authors. These free-text comments are presented to illuminate the quantitative findings. The following sections provide a summary of the key results. Appendix 5.13 provides a tabulated breakdown of the authors’ responses to each of the survey questions. As shown in Figure A1, most of the respondents to the survey were authors who had completed at least two empirical DCEs with 83% (n=44) of respondents, and 17% (n=9) having completed at least 10 DCEs.
Figure A1: Number of DCEs published as first author or co-author

Figure A2: The contribution of qualitative research methods in a health DCE study

Figure A2 shows that all the respondents perceived that using qualitative research methods had added value to the DCE they published. The authors also reported that the use of qualitative research methods added value to their experience of conducting DCEs in general, with 74% (n=31) stating it made a ‘substantial improvement’ to the study.

One respondent offered this comment, which suggests some antipathy towards the use of qualitative research methods in the DCE:

“Qualitative methods often require a subjective component that doesn't fit well with economics or quantitative methods. I am not convinced that qualitative work is always needed” (Author ID40; published 5-9 DCEs)

A key finding of the systematic review was poor reporting of qualitative research. The majority of survey respondents (n=42, 79%) agreed with this finding that the qualitative component was only briefly described in their DCE paper. Half of the respondents (n=26, 50%) believed the amount of qualitative work conducted was accurately reflected in the published paper as reported.

Figure A3 summarises the response of authors when explicitly asked about the reasons
behind the key finding of the systematic review (presented in Appendix 5.1) that qualitative research is not well reported in DCE papers. Almost half of authors (n=22, 44%) said they did not think the detail on qualitative research methods was of interest to journals:

“The subjects were asked WHY they made that choice. This yielded wonderful insight into the beliefs and misconceptions behind prevention choices. I haven't found a journal that is interested in publishing these more descriptive results” (Author ID22; published 3 DCEs)

“I have suggested to editors to incorporate a material section in the webpage including transcripts of the focus groups, and in general all the qualitative work, plus data and econometric model. This would allow replicability and learn more about real advantages and disadvantages of the methodology. Congrats for choosing such a relevant research topic and hopefully your efforts will improve the journal editorial process.” (Author ID36; published 2 DCEs)

One respondent suggested that reporting the qualitative research could jeopardise the acceptance of their paper:

“Reporting these details could cause criticism by the reviewers, and compromise acceptance of publication.” (Author ID27; published 5-9 DCEs)

Some respondents (n=11, 16%) stated that qualitative research would not be of interest to their peers. For example:

“Not all the work that was conducted is reported in the paper because of length limitations imposed by the Journal and because of the complexity of the topic to the main audience.” (Author ID27; published 5-9 DCEs)

Some (n=9, 17%) of the survey respondents also used the free-text space to comment that the journal word limit was too low to accurately incorporate the qualitative work that was conducted as this comment illustrates:

“Journal space limitations made [it] impossible to report the extensive qualitative work we had to do in order to come up with a plausible valuation scenario, attributes definition and levels.” (Author ID36; published 2 DCEs)

One author suggested that there has been too much focus on quantitative aspects of DCEs. For example:

“Perhaps there has been so much focus on the technical aspects of performing a DCE (design and analysis) that the most important step: identifying and reporting the methods for selecting attributes and their levels in a consistent manner has received less attention.” (Author ID11; published 3 DCEs)

Figure A3 shows that over half of the respondents (n=26, 52%) considered qualitative
research was too complicated to report in detail and one fifth of the respondents (n=10, 20%) reported that they felt it was too time consuming to conduct properly.

Figure A3: Reasons for limited reporting of qualitative research methods

Two authors hypothesised that the quality of the qualitative research conducted alongside DCEs meant the work did not merit extensive reporting:

“Journals usually do not allow for much reporting of this, and the qualitative studies are not necessarily conducted in a sufficient stringent approach to be reported in separate papers.” (Author ID3; published 5-9 DCEs)

“What is presented in a manuscript often is limited because of word limits or because some of the qualitative research is somewhat informal.” (Author ID19; published 10+ DCEs)

Some respondents (n=4, 8%) also reported that they did not believe qualitative research was of importance to funders. For example:

“It may not be appreciated by funders as a necessary component of doing high quality DCE studies.” (Author ID1; published 5-9 DCEs)

Three quarters of the respondents (n=40, 75%) stated that they had no expertise in qualitative research methods. Some respondents (n=31, 58%) did have a qualitative researcher as part of their team, but others (n=8, 15%) did not.

In the free-text comments, one author stated there was a lack of guidance on reporting standards for the qualitative research conducted alongside a DCE:

“Helping to disseminate guidelines/suggestions useful to report accurately but concisely details on qualitative research, in a way that it is accepted by the reviewers/readers, could help improve the transparency of the studies, hence their quality and results reliability.” (Author ID27; published 5-9 DCEs)
Discussion

A variety of reasons for the lack of reporting was identified. One respondent hypothesised that the lack of qualitative research was because it was deemed unimportant in terms of answering the research question posed. Another potential reason is lack of expertise because research teams conducting DCEs have quantitative economic backgrounds. This lack of expertise could be remedied with comprehensive guidelines on conducting qualitative research alongside a DCE study. Currently, detailed guidelines on the use of qualitative research methods only exist for the identification of attributes and levels.

The survey to authors was completed by researchers who had regular experience of designing and/or analysing DCEs, with most of the respondents publishing more than one DCE. This indicates the sample was knowledgeable and was likely to include experienced researchers whose views are likely to be representative of the wider research area. The survey found that all authors reported that qualitative research methods played a valuable role in the progression of their DCE study. This survey provided evidence that researchers designing and analysing DCEs regarded using qualitative research methods as beneficial in a health DCE study. The lack of reporting of a beneficial and informative component to a research study could be rectified with greater use of online appendices for reporting the qualitative research, particularly in word-restrictive journals.

Comparison with existing literature

None of the previous systematic reviews (Ryan & Gerard, 2003b; de Bekker-Grob et al., 2012; Clark et al., 2014) directly contacted authors to determine whether the results detailed in their papers were subject to reporting-bias. The results of the previous reviews were therefore unconditionally reliant on the textual accounts of the author as recorded in the article. Using a follow-up survey using the authors of published studies has allowed a thorough exploration of current and past DCE practice based on their own opinions.

Limitations

The survey used for this study (Appendix 5.10) had a number of limitations. Arguably, a more in-depth account of authors’ views and experiences could have been collected (possibly through one-to-one interviews) and thorough thematic analysis of the free-text comments would have provided more robust results. However, the results of the questionnaire helped to explain the key findings of the systematic review, such as the drivers behind the absence of detail, and it is unlikely further analysis or review would have highlighted anything that would significantly alter the findings.
Conclusion
The results of the systematic review and survey to authors identified qualitative research methods were being used by DCE researchers to answer multiple research questions, and that these methods add value to a DCE study. This finding suggested that qualitative research would be an appropriate method to use alongside DCEs.

Bibliography


Appendix 5.3: Flow of papers through the systematic review

Figure A1: Search results from Medline database for systematic review of studies using qualitative research alongside DCEs, run on the 30\textsuperscript{th} June 2012

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete Choice Experiment(s)</td>
<td>253</td>
</tr>
<tr>
<td>Discrete Choice Modeling</td>
<td>18</td>
</tr>
<tr>
<td>Stated Preference</td>
<td>143</td>
</tr>
<tr>
<td>Part-Worth Utilities</td>
<td>4</td>
</tr>
<tr>
<td>Functional Measurement</td>
<td>0</td>
</tr>
<tr>
<td>Paired Comparisons</td>
<td>710</td>
</tr>
<tr>
<td>Pairwise Choices</td>
<td>8</td>
</tr>
<tr>
<td>Conjoint Analysis</td>
<td>283</td>
</tr>
<tr>
<td>Conjoint Measurement</td>
<td>43</td>
</tr>
<tr>
<td>Conjoint Studies</td>
<td>2</td>
</tr>
<tr>
<td>Conjoint experiment(s)</td>
<td>2</td>
</tr>
<tr>
<td>TOTAL HITS</td>
<td>1403</td>
</tr>
<tr>
<td>Published pre-2008</td>
<td>902</td>
</tr>
<tr>
<td>RETRIEVED ABSTRACTS</td>
<td>501</td>
</tr>
</tbody>
</table>
Figure A2: Detailed reasons for inclusion and rejection
### Appendix 5.4: Comparison of existing appraisal tools for assessing the quality of qualitative research

#### Table A1: Comparison of overview and context criteria covered by existing appraisal tools

<table>
<thead>
<tr>
<th>Clear aims or hypothesis</th>
<th>Importance and relevance of research</th>
<th>Appropriateness of qualitative methodology</th>
<th>Appropriate research design for aims?</th>
<th>Details of phenomena being studied</th>
<th>Thorough and systematic literature review to link research to current knowledge</th>
<th>Theoretical framework for study</th>
<th>Study perspectives (service, user carer etc)</th>
<th>Discussion of time horizon of study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Appraisal Skills Programme (CASP)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Prompts for appraising qualitative research (Dixon-Woods)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation Tool for Qualitative Studies (Long)</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Quality Framework (Focus)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale and Standards (Popay)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appraising the Quality of Qualitative Research (Walsh)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JBI QARI</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Table A2: Comparison of sampling and data collection criteria covered by existing appraisal tools

<table>
<thead>
<tr>
<th></th>
<th>Discussion of recruitment strategy and technique (purposive, convenience etc)</th>
<th>Key characteristics of sample</th>
<th>Why these people were selected</th>
<th>Details of comparison groups</th>
<th>Appropriateness of setting</th>
<th>Clear description of data collection method (to replicate)</th>
<th>Justified collection methods</th>
<th>Explicit description of process (topic guides etc)</th>
<th>Reporting of modification of methods in study</th>
<th>Form of data (tape records)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Appraisal Skills Programme (CASP)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prompts for appraising qualitative research (Dixon-Woods)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation Tool for Qualitative Studies (Long)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality Framework (Focus)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rationale and Standards (Popay)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appraising the Quality of Qualitative Research (Walsh)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JBI QARI</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table A3: Comparison of reflexivity and ethics criteria covered by existing appraisal tools

<table>
<thead>
<tr>
<th>Researcher(s) have discussed their own potential bias</th>
<th>Study demonstrates that ethical standards were maintained/ issues addressed</th>
<th>Handling of data and respondents post/during study</th>
<th>Approval from committee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Appraisal Skills Programme (CASP)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Prompts for appraising qualitative research (Dixon-Woods)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation Tool for Qualitative Studies (Long)</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Quality Framework (Focus)</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Rationale and Standards (Popay)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appraising the Quality of Qualitative Research (Walsh)</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>JBI QARI</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>Description of analysis process</td>
<td>Derivation of themes discussed</td>
<td>Evidence of iterative analysis</td>
<td>Steps to guard against selectivity</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>--------------------------------</td>
<td>-----------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Critical Appraisal Skills Programme (CASP)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Prompts for appraising qualitative research (Dixon-Woods)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Evaluation Tool for Qualitative Studies (Long)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Quality Framework (Focus)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Rationale and Standards (Popay)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Appraising the Quality of Qualitative Research (Walsh)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JBI QARI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
**Table A5: Comparison of results and value of research criteria covered by existing appraisal tools**

<table>
<thead>
<tr>
<th></th>
<th>Explicit stating of the findings</th>
<th>Adequate discussion of evidence for/against</th>
<th>Credibility and validity demonstrated (triangulation etc)</th>
<th>Findings in relation to research question</th>
<th>Comparison to other studies</th>
<th>Contribution to existing literature</th>
<th>Recommendation for future research</th>
<th>Generalisability of results</th>
<th>Policy and practice implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Appraisal Skills Programme (CASP)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Prompts for appraising qualitative research (Dixon-Woods)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation Tool for Qualitative Studies (Long)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Quality Framework (Focus) Rationale and Standards (Popay)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Appraising the Quality of Qualitative Research (Walsh)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>JBI QARI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 5.5: Final critical appraisal tool for assessing the quality of qualitative research alongside DCEs

<table>
<thead>
<tr>
<th>A CHECKLIST FOR QUALITATIVE RESEARCH ALONGSIDE STATED PREFERENCE METHODS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OVERVIEW</strong></td>
</tr>
<tr>
<td>Are there clear aims, objectives or hypotheses to the qualitative component?</td>
</tr>
<tr>
<td>Does the research fill a gap or link to current knowledge?</td>
</tr>
<tr>
<td><strong>CONTEXT</strong></td>
</tr>
<tr>
<td>Is there a theoretical framework/approach for the study? (Grounded theory, phenomenology, etc.)?</td>
</tr>
<tr>
<td>Is there a clear application for the study? Does the setting match the aims?</td>
</tr>
<tr>
<td>Is it clear and justified why the qualitative data was collected concurrently or retrospectively?</td>
</tr>
<tr>
<td><strong>SAMPLING</strong></td>
</tr>
<tr>
<td>Is there a clear description of the recruitment strategy?</td>
</tr>
<tr>
<td>Are the details of the sample’s characteristics provided?</td>
</tr>
<tr>
<td>Does the sample contribute to the generalisability of the study? Did the recruitment result in an unbiased sample?</td>
</tr>
<tr>
<td>Does the study consider the appropriateness of the selected sample? Will this match the final or typical stated preference survey respondents?</td>
</tr>
<tr>
<td>Did sampling continue until data saturation?</td>
</tr>
<tr>
<td>Is the response rate detailed? Is there response rate poor? Why?</td>
</tr>
<tr>
<td><strong>COLLECTION</strong></td>
</tr>
<tr>
<td>Are adequate descriptions of the methods and process for data collection? Could the study be replicated?</td>
</tr>
<tr>
<td>Are the use of qualitative methods appropriate and justified? Are the schedules or prompts presented in, or with, the paper?</td>
</tr>
<tr>
<td>Are examples of the survey (DCE or CV) questions presented? How were preferences elicited?</td>
</tr>
<tr>
<td>Is there a clear description of the form of data? (tape recordings, videos, free-text comments)</td>
</tr>
<tr>
<td><strong>REFLEXIVITY</strong></td>
</tr>
<tr>
<td>Is there a description of the qualifications or experience of the researchers involved?</td>
</tr>
<tr>
<td><strong>ETHICS</strong></td>
</tr>
<tr>
<td>Is there an indication of ethical approval if the qualitative study is dealing with a sensitive subject?</td>
</tr>
<tr>
<td><strong>ANALYSIS</strong></td>
</tr>
<tr>
<td>Is the analytical process adequately described? If themes were identified, how did this happen?</td>
</tr>
<tr>
<td>Is there evidence of triangulation? Were other sources involved in the interpretation?</td>
</tr>
<tr>
<td>Are raw data, such as verbatim quotes, presented or made available on request?</td>
</tr>
<tr>
<td>Is there an explanation of why the presented data were selected?</td>
</tr>
<tr>
<td>Is there congruity between the presented data and the interpretations drawn?</td>
</tr>
<tr>
<td><strong>RESULTS</strong></td>
</tr>
<tr>
<td>Are the findings of the qualitative research clearly stated?</td>
</tr>
<tr>
<td>Has the paper considered the credibility and validity of results? Are there reports of reflexivity?</td>
</tr>
<tr>
<td>Are conflicting qualitative results explored?</td>
</tr>
<tr>
<td>Do the qualitative results relate to the original research question(s)?</td>
</tr>
<tr>
<td>Is there a comparison of the qualitative results to existing literature or similar studies?</td>
</tr>
<tr>
<td>Are recommendations for future qualitative or quantitative research made?</td>
</tr>
<tr>
<td>Is the generalisability of the qualitative results discussed?</td>
</tr>
<tr>
<td>Is there a clear application for the results of the qualitative research? (Will it change or add to attributes and levels in a DCE? Does this inform the starting bid for a willingness-to-pay study?)</td>
</tr>
</tbody>
</table>
Appendix 5.6: Studies included in the systematic review of qualitative research conducted alongside DCEs

<table>
<thead>
<tr>
<th>ID</th>
<th>Full reference</th>
</tr>
</thead>
</table>


Table A1: Number and percentage of studies included in the systematic review by year of publication

<table>
<thead>
<tr>
<th>Year</th>
<th>n</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>7</td>
<td>(3%)</td>
</tr>
<tr>
<td>2002</td>
<td>8</td>
<td>(3%)</td>
</tr>
<tr>
<td>2003</td>
<td>7</td>
<td>(3%)</td>
</tr>
<tr>
<td>2004</td>
<td>12</td>
<td>(4%)</td>
</tr>
<tr>
<td>2005</td>
<td>12</td>
<td>(4%)</td>
</tr>
<tr>
<td>2006</td>
<td>20</td>
<td>(7%)</td>
</tr>
<tr>
<td>2007</td>
<td>27</td>
<td>(10%)</td>
</tr>
<tr>
<td>2008</td>
<td>25</td>
<td>(9%)</td>
</tr>
<tr>
<td>2009</td>
<td>41</td>
<td>(15%)</td>
</tr>
<tr>
<td>2010</td>
<td>42</td>
<td>(16%)</td>
</tr>
<tr>
<td>2011</td>
<td>58</td>
<td>(21%)</td>
</tr>
<tr>
<td>2012*</td>
<td>11</td>
<td>(4%)</td>
</tr>
</tbody>
</table>

*2012 up to search date

Table A2: Number and percentage of studies included in the systematic review by country

<table>
<thead>
<tr>
<th>Country</th>
<th>n</th>
<th>%</th>
<th>Country</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>23</td>
<td>8.5%</td>
<td>Multi Country</td>
<td>20</td>
<td>7.4%</td>
</tr>
<tr>
<td>Canada</td>
<td>16</td>
<td>5.9%</td>
<td>Nepal</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>Denmark</td>
<td>9</td>
<td>3.3%</td>
<td>Netherlands</td>
<td>33</td>
<td>12.2%</td>
</tr>
<tr>
<td>Finland</td>
<td>1</td>
<td>0.4%</td>
<td>Russia</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>France</td>
<td>3</td>
<td>1.1%</td>
<td>South Africa</td>
<td>2</td>
<td>0.7%</td>
</tr>
<tr>
<td>Germany</td>
<td>7</td>
<td>2.6%</td>
<td>South Korea</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>Ghana</td>
<td>3</td>
<td>1.1%</td>
<td>Spain</td>
<td>4</td>
<td>1.5%</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>2</td>
<td>0.7%</td>
<td>Sweden</td>
<td>2</td>
<td>0.7%</td>
</tr>
<tr>
<td>Hungary</td>
<td>1</td>
<td>0.4%</td>
<td>Switzerland</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>Ireland</td>
<td>1</td>
<td>0.4%</td>
<td>Taiwan</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>Italy</td>
<td>5</td>
<td>1.8%</td>
<td>Tanzania</td>
<td>2</td>
<td>0.7%</td>
</tr>
<tr>
<td>Ivory Coast</td>
<td>1</td>
<td>0.4%</td>
<td>Thailand</td>
<td>1</td>
<td>0.4%</td>
</tr>
<tr>
<td>Japan</td>
<td>4</td>
<td>1.5%</td>
<td>UK</td>
<td>76</td>
<td>28.0%</td>
</tr>
<tr>
<td>Liberia</td>
<td>1</td>
<td>0.4%</td>
<td>USA</td>
<td>43</td>
<td>15.9%</td>
</tr>
<tr>
<td>Malawi</td>
<td>1</td>
<td>0.4%</td>
<td>Vietnam</td>
<td>2</td>
<td>0.7%</td>
</tr>
<tr>
<td>Missing</td>
<td>2</td>
<td>0.7%</td>
<td>Zambia</td>
<td>1</td>
<td>0.4%</td>
</tr>
</tbody>
</table>
Table A3: Number and percentage of studies included in the systematic review by journal type

<table>
<thead>
<tr>
<th>Journal Type</th>
<th>n</th>
<th>% of sub discipline</th>
<th>% of total studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health Services Research</td>
<td>1</td>
<td>-</td>
<td>51%</td>
</tr>
<tr>
<td>Health Economics</td>
<td>7</td>
<td>57%</td>
<td>29%</td>
</tr>
<tr>
<td>Health services research and health policy</td>
<td>6</td>
<td>43%</td>
<td>22%</td>
</tr>
<tr>
<td>journals</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specialist medicine</td>
<td>1</td>
<td>-</td>
<td>38%</td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Cardiology</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Dentistry</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Dermatology</td>
<td>5</td>
<td>5%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Diet</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Endocrinology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Genetics</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Geriatrics</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Hematology</td>
<td>4</td>
<td>4%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Immunology</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Nephrology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Neurology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Obstetrics and gynecology</td>
<td>1</td>
<td>13%</td>
<td>4.7%</td>
</tr>
<tr>
<td>Oncology</td>
<td>9</td>
<td>9%</td>
<td>3.2%</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Orthopedics</td>
<td>3</td>
<td>3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Physical medicine and rehabilitation</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Psychiatry</td>
<td>3</td>
<td>3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Psychology</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Reproductive medicine</td>
<td>4</td>
<td>4%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Respiratory medicine</td>
<td>1</td>
<td>13%</td>
<td>4.7%</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>3</td>
<td>3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Surgery</td>
<td>2</td>
<td>2%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Urology</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Venerology</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Telemedicine</td>
<td>3</td>
<td>3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Emergency Care</td>
<td>3</td>
<td>3%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Nursing</td>
<td>1</td>
<td>1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Pharmacology</td>
<td>1</td>
<td>10%</td>
<td>3.6%</td>
</tr>
<tr>
<td>General Medicine</td>
<td>3</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 5.8: List of studies not included in the Clark et al. (2014) systematic review


Appendix 5.9: The submitted paper describing the development of the appraisal tool

CHEQUAL: a checklist for the use of qualitative methods in stated preference valuation studies

<table>
<thead>
<tr>
<th>Journal</th>
<th>Health Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>HEC-15-0445</td>
</tr>
<tr>
<td>Wiley - Manuscript type</td>
<td>Health Economics Letter</td>
</tr>
<tr>
<td>Keywords</td>
<td>Qualitative research, stated preference, discrete choice experiment, contingent valuation method, checklist</td>
</tr>
</tbody>
</table>

http://mc.manuscriptcentral.com/hec
Title: CHEQUAL: a checklist for the use of qualitative methods in stated preference valuation studies

Key words: Qualitative research, stated preference, discrete choice experiment, contingent valuation method, critical appraisal, checklist
Summary [200 words]

Qualitative research methods can inform the design and interpretation of stated preference valuation studies. The potential strengths of qualitative research can only be realised if studies are conducted appropriately and reported with clarity. This study had three aims: to identify whether existing qualitative research critical appraisal tools are suitable for use regarding qualitative research conducted as part of a stated preference valuation study, make recommendations on the modification of existing tools to produce a new checklist; and test the proposed checklist in an exemplar systematic review of stated preference studies. An iterative process of identifying, testing and critiquing existing appraisal tools was used to produce a checklist for use when reporting qualitative methods used alongside a stated preference valuation study. Seven established qualitative appraisal tools were identified in the health services research literature. However, these tools were judged to be inappropriate for use in respect to stated preference valuation studies. A checklist was designed for health economists using qualitative methods in stated preference valuation studies ('CHEQUAL'). The new checklist was tested in a systematic review of healthcare discrete choice experiments. CHEQUAL offers a transparent mechanism for extracting, summarising and critiquing the reported qualitative research used in stated preference valuation studies.
I. Introduction

The need to promote rigour in published research has stimulated the development of checklists for the appraisal of study designs used in health economics (for example, the Consolidated Health Economic Evaluation Reporting Standards (Husereau et al., 2013; Xie et al., 2015). These checklists form the basis for appraisal tools that can be used to inform the design of studies and preparation of manuscripts for submission to journals or the critical appraisal of published studies in the context of systematic reviews. Similarly, peer-reviewers, funders and decision-makers, interested in assessing or using the research, can use these checklists to provide a structured view on the quality of published studies as reported.

‘Qualitative research’ is increasingly advocated in the field of health economics (Coast, 1999; Coast et al., 2004). The term ‘qualitative research’ refers to a broad range of philosophies, approaches and methodologies used to acquire an in-depth understanding or explanation of people’s perceptions (Braithwaite & Clarke, 2006; Corbin & Strauss, 2008; Hsieh & Shannon, 2005; Pope & Mays, 2008). Some commentators have made specific recommendations for application of qualitative research in the design and interpretation of quantitative stated preference (SP) studies (see figure 1 for an overview of types of SP studies) (Coast et al., 2012; Klijn et al., 2012; Baker & Robinson, 2004; Baker et al., 2008). A key strength of qualitative methods is being able to collate important contextual, alongside quantitative preference, data (Cheraghi-Sohb et al., 2007; Ryan et al., 2009). The potential strengths of qualitative research can only be realised if studies are conducted appropriately and reported with sufficient clarity to understand the methods and interpretation of the findings.

<Figure 1 here>
Quantitative SP studies are survey-based valuation methods increasingly used in health services research to elicit individuals' preferences for goods or services (Clark et al., 2014; de Bekker-Grob et al., 2012; Smith & Sach, 2010). Guidelines for the design and conduct of SP studies recommend the use of qualitative methods (Bridges et al., 2011; Lancsar & Louviere, 2008). Existing guidelines have not provided a checklist on how to report the use of qualitative research alongside SP valuation studies. This study had three objectives to: identify whether existing qualitative critical appraisal tools are suitable for use with SP studies, if appropriate, make recommendations on how existing critical appraisal tools could be modified appropriately to produce a checklist, and, test the feasibility of using the suggested checklist in an exemplar systematic review of published healthcare SP valuation studies.

2. Methods

Existing qualitative methods critical appraisal tools (hereafter termed ‘tools’) were identified from guidelines published by the Centre for Reviews and Dissemination (CRD) (CRD, 2008). Manual searching of qualitative websites was combined with personal communication with two researchers with expertise in conducting qualitative healthcare studies. A tool was included if it was designed for appraising qualitative research generically, rather than for specific methods (e.g. ethnographic studies, interview studies).

The identified tools were summarised (independently by researchers XXX and XX) in a tabular format to facilitate direct comparison of the suggested quality assessment criteria and coverage of the existing tools. In addition, two published SP studies known to include a substantial qualitative aspect (Cheraghi-Sohi et al., 2007; Ryan et al., 2009) were used.
(independently by two researchers: XXX and XX) to test the identified tools that offered checklists to assess the quality of published qualitative studies.

Discussions were then held between members of the research team (with primary expertise in: health economics (n=3); economics (n=1); qualitative research (n=2)) and informed by the tabulated summary criteria with the exemplar use of data extracted using the identified tools. This iterative process aimed to reach consensus on the relevance of each tool to inform a checklist of the qualitative elements of SP studies. The process identified that important criteria were absent but some included criteria were afforded an unwarranted importance; consequently concluding that no existing tool adequately met the requirements for the purpose of assessing qualitative research used alongside SP studies. Discussion of existing tools produced a new set of criteria to be included in a preliminary version of a checklist. The feasibility and practical use of the proposed checklist was then tested in two published DCEs (Ryan et al., (2009) and Cheraghi-Sohi et al., (2007)) and two contingent valuation studies (Smith, (2007) and Shiell & Gold (2002)) all of which reported the use of qualitative research methods in either the survey design or interpretation of the results. Data were extracted from articles in accordance with the proposed checklist and further discussions were held between research team members to review how capable the checklist was of capturing the dimensions of the qualitative research conducted. This iterative process of testing, refining and discussing continued until an acceptable list of criteria was reached. The final version of the checklist was then agreed and further tested in a systematic review of qualitative research methods used alongside published healthcare DCEs.

3. Results

http://mc.manuscriptcentral.com/hec
Seven published qualitative methods critical appraisal tools (‘tools’) were identified and reviewed (See Appendix A). The tools covered eight broad themes: 1) background and overview; 2) description of the context; 3) sampling methods; 4) data collection; 5) reflexivity (or a consideration of how the investigators themselves can influence qualitative research); 6) ethical aspects; 7) data analysis and 8) presentation of results. Across these eight themes a set of 32 specific criteria were identified. The number of criteria included in the tools ranged from seven to 42.

There was agreement (amongst all members of the research team) that none of the existing tools were appropriate to inform a checklist of qualitative methods in SP valuation studies for three key reasons that centred on the rationale for using qualitative methods; reporting of philosophical foundations and theoretical underpinnings of the selected qualitative method; and practical use of the checklist. Certain aspects of some of tools were identified to be potentially relevant and served as the starting point for a new checklist.

The available tools did not allow an assessment of whether a clear rationale for using qualitative methods was reported in the SP study. This is an important omission regarding qualitative methods’ use with SP valuation studies. It should be explicit whether qualitative research methods were employed in: 1) the design phase and/or 2) order to interrogate the validity of the quantitative data. These two rationale, respectively, have distinct implications for the study findings, as qualitative methods should have been used and reported with sufficient clarity to understand whether: 1) the data had been reduced to generate a final list of attributes and levels that were framed and presented in a way that respondents could understand; and 2) the results were in accordance with underlying economic theory and

http://mc.manuscriptcentral.com/nec
rationale for the use of choice-based studies and the impact on model selection, econometric tests or assumptions.

One of the existing tools placed substantial emphasis on the need to report the philosophical foundations of using qualitative research methods (Long & Godfrey, 2004). Similarly, two of the tools (Long & Godfrey (2004) and Popay et al. (1998)) placed emphasis on reporting: the theoretical framework; the epistemological and ontological foundations; and evidence of conceptual adequacy. Details of these criteria may not be reported in a SP valuation study due to journal word restrictions and/or a perception that an economic audience is not sufficiently familiar or interested in these aspects. It was unanimously agreed that the omission of these aspects would not detract from an assessment of the quality of the qualitative research in the context of SP valuation studies.

In the 1990s, a consistent view emerged that mixed methods (qualitative combined with quantitative) studies in health services research would add value to knowledge and understanding (Barbour, 1999; Pope & Mays, 1993, 1995; O'Cathain et al., 2007). However, the use of mixed methods studies did not result in the desired effect, with criticism of each component, either in isolation or when combined, due to inadequate design and/or reporting (O'Cathain et al., 2008; Pope & Mays, 2009). The proposed new checklist, unlike existing tools, was designed to determine if, and how, qualitative and quantitative data were synthesised and included in the study design and/or interpretation.
Two further issues with using existing tools were identified. The tools failed to take into account the style of studies published in medical or health economics journals and the tension between reporting the mixed methods employed within tight word limits. The restriction of word count can, in part, be addressed by judicious use of supplementary appendices, but clarity is still required on methods and analysis. Another practical issue is that most health economists, who design SP valuation studies, are not trained in the use of qualitative methods. This is often addressed by working with qualitative researchers but then another tension emerges which is whether the study report should be focussed on the qualitative or quantitative aspects. The proposed checklist is designed to provide a means of rigorous peer-review of the qualitative component alongside the quantitative aspects of a SP valuation study by a non-expert user.

The iterative process produced a new checklist called ‘CHEQUAL’; a mechanism to facilitate the systematic reporting and evaluation of the use of qualitative methods in SP valuation studies (see Appendix B). CHEQUAL could also guide researchers seeking to design a qualitative component of a SP study.

The final version of CHEQUAL was tested by using it to summarise if, and how, qualitative research methods were reported in published healthcare DCEs, generating a tabular summary using the 32 identified criteria. All healthcare DCEs published between 2001 and 2012 were identified (n=254) using an electronic search strategy based on de Bekker-Grob et al (2012). Of the 254 DCEs, 143 studies explicitly reported some use of qualitative methods and of these, 29 reported an extensive qualitative component either in the article or as a citation to previous work. A study was defined as having an extensive qualitative component (either in

http://mc.manuscriptcentral.com/hec
the article or in cited related work) when three conditions were met: a clear statement of the purpose of the qualitative research; an indication of the technique used to collect qualitative data; a description of how the qualitative data was examined/software used; and an explanation of the outcomes of the qualitative research. The feasibility of using CHEQUAL was demonstrated and the checklist was judged to identify key pieces of information which were, and were not, reported.

4. Discussion

The use of qualitative techniques to inform the design or understanding of SP studies is a relatively new but increasingly popular methodological direction within health economics. This study identified a need for a checklist designed specifically to check the extent to which, and how, qualitative research was conducted alongside SP studies. A practical and feasible checklist (CHEQUAL) was developed and tested. It is anticipated that this checklist will be of use to health economists, and their qualitative research co-investigators, to aid the clear and transparent reporting of the design, analysis and use of qualitative methods used in existing SP valuation studies. It could also prove a useful starting point for researchers embarking on a new SP study to inform the design of the qualitative component of the study. The proposed checklist also has a role in the conduct of systematic reviews of published SP valuation studies to critically appraise how qualitative methods were used alongside the quantitative aspects.

5. Conclusion

The development of CHEQUAL is timely given the increased use of qualitative research methods alongside SP valuation studies. The proposed new checklist was designed specifically to offer a transparent, fair and balanced mechanism for standardising the
reporting of qualitative research alongside SP valuation studies. SP valuation studies have a
total role to play in health services research and as a source of information for resource
allocation decision-making. The benefits of using qualitative methods alongside quantitative
methods can only be realised if details are reported with clarity. CHEQUAL has been pre-
tested in the context of healthcare DCIs but to meet multiple research and policy objectives
the checklist should be further tested in wider application in the field of health economics.
Acknowledgements

We would like to acknowledge the discussions with colleagues who attended a seminar that presented the potential use of CHEQUAL at the Health Economics Unit, University of Birmingham. In particular, we want to thank Dr Emma Frew for her considered suggestions at this seminar.

Declaration: The authors have no financial conflicts of interests. No ethical approval was required for the research conducted in this study.
References


http://mc.manuscriptcentral.com/hec


http://mc.manuscriptcentral.com/hec


Appendix 5.10: The questionnaire sent to authors of DCE studies to assess their experience and views of using qualitative research methods

The role and reporting of qualitative research methods in DCEs: a survey of study authors

Thank you for considering taking part in this survey.

The research is part of a PhD project at the University of Manchester which is sponsored by the National Institute of Health Research’s School for Primary Care Research.

You have been invited to participate because you have published at least one discrete choice experiment (DCE) which has been included in my systematic review.

The survey will take approximately 10 minutes to complete.

To enter the survey, press the forward arrow below.

If you want to speak to someone about the research or this survey, please contact Caroline Vass, caroline.vass@postgrad.manchester.ac.uk or +44 (0) 161 306 7970

Thank you for agreeing to complete this survey.

The purpose of this survey is to identify your views on the role and reporting of Qualitative Research Methods (QRM) in

(1) Designing a discrete choice experiment (DCE) and/or

(2) Understanding the responses to a DCE.
How many DCEs in health care have you published, either as the first author or as a co-author?

- 1
- 2
- 3
- 4
- 5 - 9
- 10 or more

National School for Primary Care Research

For this question, please think about the study mentioned in the invitation email.

Do you think the paper accurately reflected the amount of qualitative research undertaken in the study?

- Yes
- No
- Don't know

Would you like to expand on your response?
For this question, please think about the study mentioned in the invitation email.

Was a member of the research team an expert in qualitative research?

Regarding expertise in qualitative research...

If no, could you please briefly explain your answer.

A key finding of my systematic review of DCEs in health care was that many studies report either limited or no qualitative research methods.

Does this finding agree with your experience of reading or conducting health care DCEs?

- Yes
- No (please explain why)
Why do you think there is this apparent absent / limited reporting of qualitative research methods? Some possible reasons are listed below, please tick all which apply

- Qualitative research in DCEs is not of interest to most of my peers
- Qualitative research in DCEs is not of interest to journals
- Qualitative research in DCEs is not of interest to funders
- Qualitative research is not important in the design of health DCEs
- Qualitative research does not affect the study outcomes
- Qualitative research is too complicated to report in detail
- Qualitative research is too time consuming to conduct properly
- Other reasons (please describe)

---

Do you feel that the qualitative research completed as part of this health care DCE added value?

- It made a substantial improvement
- It added a little value
- None, it did not add any value at all
- No, it hindered the study and had a negative role
Thinking about the study mentioned in the invitation email, do you feel that the qualitative research completed as part of your DCE added value?

- It made a substantial improvement
- It added a little value
- None, it did not add any value at all
- No, it hindered the study and had a negative role

National School for Primary Care Research

Thinking about all DCEs you have conducted, in general, how would you describe the impact of the qualitative research?

- It makes a substantial improvement
- It adds a little value
- It has no impact at all
- It has a negative impact and hinders the study progression

National School for Primary Care Research
Is there anything you would like to add regarding the issues this survey has addressed or the questions we have asked you?

National School for Primary Care Research

Would you like a copy of the results of this review?

- Yes
- No

National School for Primary Care Research
To receive a copy of the results, please provide your email address in the space below.

The survey has now been submitted. Thank you.

For any queries or further information, please contact Caroline Vass (caroline.vass@postgrad.manchester.ac.uk)
Appendix 5.11: The invitation to authors of DCE studies to complete the questionnaire to assess their experience and views of qualitative research methods alongside DCEs

Dear <TitleA> <Last Name>,

I am a PhD student sponsored by the National Institute for Health Research (NIHR) National School for Primary Care Research and based in the Manchester Centre for Health Economics at The University of Manchester, UK. I am supervised by Professor Katherine Payne looking at the framing of risk attributes in discrete choice experiments (DCEs).

I am writing to ask if you would complete an on-line survey that should take approximately 5 minutes of your time. The purpose of this survey is identify your views on the role and reporting of qualitative research methods in DCEs.

I have conducted a systematic review of the use of qualitative research methods in published DCEs. A summary of this systematic review is shown at the end of this email. A limitation of this systematic review is that it can only summarise the reported qualitative research methods. This is why I am contacting you, as an author of a published DCE, to complete this short survey. I want to understand more about researchers’ views of using qualitative methods in DCEs with particular reference to your paper: <Title> in <Journal> (<Year>).

To complete the survey please click on this link <link>

Please note that your responses to this survey will be anonymously reported as a summary of the key findings.

Thank you very much for taking the time to read this email and I hope you are able to complete my survey for me.

Kind regards,

Caroline Vass

Manchester Centre for Health Economics
NIHR School for Primary Care Research Doctoral Researcher
Institute of Population Health, Jean McFarlane Building
The University of Manchester
Oxford Road, Manchester M13 9PL
Appendix 5.12: The reminder invitation to authors of DCE studies to complete the questionnaire to assess their experience and views of qualitative research methods alongside DCEs

Dear <TitleA> <Last Name>,

This is a friendly reminder that the invitation to complete this survey is still open. Your response would be greatly appreciated.

I am a PhD student sponsored by the National Institute for Health Research (NIHR) National School for Primary Care Research and based in the Manchester Centre for Health Economics at The University of Manchester, UK. I am supervised by Professor Katherine Payne looking at the framing of risk attributes in discrete choice experiments (DCEs).

I am writing to ask if you would complete an on-line survey that should take approximately 5 minutes of your time. The purpose of this survey is identify your views on the role and reporting of qualitative research methods in DCEs.

I have conducted a systematic review of the use of qualitative research methods in published DCEs. A summary of this systematic review is shown at the end of this email. A limitation of this systematic review is that it can only summarise the reported qualitative research methods. This is why I am contacting you, as an author of a published DCE, to complete this short survey. I want to understand more about researchers’ views of using qualitative methods in DCEs with particular reference to your paper: <Title> in <Journal> (<Year>).

To complete the survey please click on this link <link>

Please note that your responses to this survey will be anonymously reported as a summary of the key findings.

Thank you very much for taking the time to read this email and I hope you are able to complete my survey for me.

Kind regards,

Caroline Vass

Manchester Centre for Health Economics
NIHR School for Primary Care Research Doctoral Researcher
Institute of Population Health, Jean McFarlane Building
The University of Manchester
Oxford Road, Manchester M13 9PL
Appendix 5.13: Author responses to the questionnaire of their experience and views of conducting qualitative research alongside DCEs

| How many DCEs in health care have you published, either as the first author or as a co-author? |
|---|---|
| n | % |
| 1 | 9 | 17% |
| 2 | 10 | 19% |
| 3 | 9 | 17% |
| 4 | 0 | 0% |
| 5-9 | 16 | 30% |
| 10+ | 9 | 17% |

| Why do you think there is this apparent absence / limited reporting of qualitative research methods? Some possible reasons are listed below, please tick all which apply. Qualitative research in DCEs... |
|---|---|
| Is not of interest to most of my peers | 11 | 22% |
| Is not of interest to journals | 22 | 44% |
| Is not of interest to funders | 4 | 8% |
| Is not important in the design of health DCEs | 3 | 6% |
| Does not affect the study outcomes | 2 | 4% |
| Is too complicated to report in detail | 26 | 52% |
| Is too time consuming to conduct properly | 10 | 20% |
| Other reasons | 28 | 56% |

| Do you feel that the qualitative research completed as part of this health care DCE added value? |
|---|---|---|
| In this DCE | In DCEs generally |
| n | % | n | % |
| It made a substantial improvement | 29 | 58% | 31 | 74% |
| It added a little value | 21 | 42% | 11 | 26% |
| None, it did not add any value at all | 0 | 0% | 0 | 0% |
| No, it hindered the study and had a negative role | 0 | 0% | 0 | 0% |

For this question, please think about the study mentioned in the invitation email. Was a member of the research team an expert in qualitative research?

| Yes, I have expertise in qualitative research | 13 | 25% |
| Yes, a member of the research team had expertise in qualitative research | 31 | 58% |
| Do not know | 1 | 2% |
| No, there was no expert in qualitative research in the research team | 8 | 15% |

Do you think the paper accurately reflected the amount of qualitative research undertaken in the study?

| Yes | 26 | 50% |
| No | 24 | 46% |
| Don’t know | 2 | 4% |

A key finding of my systematic review of DCEs in health care was that many studies report either limited or no qualitative research methods. Does this finding agree with your experience of reading or conducting health care DCEs?

| Yes | 42 | 79% |
| No | 11 | 21% |
Appendix 5.14: Paper advertisement to the general public for participation in the think-aloud interviews

Understanding risk in healthcare: an interview study

Volunteers Required
We make choices that involve risk all the time. How we travel, what we choose to eat, all involve differences in risk. Experts have created new ways of explaining differences in risk levels, and we would like to know your opinion of them.

We are looking for female volunteers aged 18-70 to take part in an interview. The interview should take no more than 45 minutes and can take place at a location convenient for you.

If you would like to find out more information, please contact Caroline Vass on the details below.

Tel 0771 675 1262
@ caroline.vass@postgrad.manchester.ac.uk

The University of Manchester
National Institute for Health Research
Appendix 5.15: Email advertisements to The University of Manchester staff for participation in the think-aloud interviews

To: Distribution Lists

Subject: Call for Research Participants

Dear Staff,

I am a PhD student in the Manchester Centre for Health Economics. This is an invitation to take part in an interview that aims to understand your views about risk in healthcare and identify how balance the risks and benefits of breast cancer screening.

The interview will take approximately 45 minutes and can take place at a location convenient for you. The full information sheet about what to expect should you choose to participate is attached.

Please feel free to forward this email to anyone else, in or outside the university, who might be interested in participating in this research.

Many thanks,

Caroline Vass

PhD Student in Health Economics
The Manchester Centre for Health Economics
Institute of Population Health
University of Manchester
4.306, Jean McFarlane Building
Oxford Road
Manchester M13 9PL
Tel: +44 (0)161 306 7970
http://www.population-health.manchester.ac.uk/students/CarolineVass
Appendix 5.16: Think-aloud interview schedule

**Introduction:** My name is Caroline Vass and I’m a PhD student at the Manchester Centre for Health Economics. My PhD is about how people make a balance between risks and benefits when making choices in health care. I am particularly interested in how we can use surveys to understand their views on risks and benefits.

This interview is about how you balance risks and benefits in health care. I will ask you to complete a questionnaire and to say out-loud how you have come to your answers.

The interview should take about 45 minutes, and there will be imaginary examples looking at breast cancer screening programmes.

Have you read the information sheet? Please complete this consent form. So, are you ok to continue?

**Topic 1: Warm-Up Exercise**

I am going to ask you a question and I would like you to think aloud as you answer it. What I mean by think aloud is basically for you to say aloud everything that you would normally say to yourself or are thinking about silently.

I know that it is not something you would normally do but it will help me understand what you are thinking and how you came up with your answer. Do you understand what I would like you to do?

**Question 1:** Would you mind telling me how many windows do you have in your house?

**Topic 2: The DCE**

Here is a survey made up of questions about imaginary breast cancer screening programmes. Take a few minutes to read through the information and definitions at the front of the survey. Do you have any questions at this point?

**Question 1:** As you complete the survey I want you to talk aloud by trying to say out-loud what you’re thinking as you’re completing this questionnaire.

Don’t worry about making sense or talking to me, I am merely going to listen to what you have to say. If you go a bit quiet, I might ask some questions to get you talking again. Is that ok?

Probes:

- What are you thinking now?
- Why did you choose that one?
- Would you choose that if I wasn’t here?
• Are you considering all of the information presented?
  [If respondent is choosing A's or B's, probe why…]
  [If respondent fails internal tests [QUESTION 6], probe why…]

**Topic 3: Ease of Completion**

Question 1: What did you think about the questions?
Question 2: Did you find yourself concentrating on a particular characteristic or do you think you weighed them up evenly?
Question 3: Would you change any of your answers?
Question 4: Do you think you are satisfied with your choices?

**Topic 4: Further questions**

These questions are just so I can understand if different people see things differently.
Question 1: What is your current employment status? What kind of work do / did you do? Was/Is that full or part-time? Do you have to use numbers in this job?
Question 2: Do you play games that involve a gamble? If not clear, suggest betting on horses or bingo. Where do you play <insert game>? Is that online?

**Topic 5: Final thoughts**

Do you have any other feedback or thoughts?
Thank you for completing this interview.
Appendix 5.17: Flow of recruited study participants for the think-aloud interviews and the eye-tracking study
Appendix 6.1: Paper advertisement to the general public for participation in the eye-tracking study

Understanding risk in healthcare: 
an eye-tracking study

Volunteers Required
We make choices that involve risk all the time – deciding how to travel and what to eat and drink can all involve choices about risk.
This is true in healthcare – for example deciding whether to undergo a treatment that could improve our health, but might involve side effects.
To help patients decide what to do, doctors sometimes provide information that helps explain different risks.

We are looking for female volunteers aged 18 to 70 to take part in a study at the University in which volunteers will be asked to look at information about breast cancer screening programmes. We'll ask their opinions about the information provided and check what information people’s eyes are drawn to. It should take no more than 25 minutes.

For more information please contact
Caroline Vass

Tel 0771 675 1262
@ caroline.vass@postgrad.manchester.ac.uk
Appendix 6.2: Email advertisement to The University of Manchester staff for participation in the eye-tracking study

To: Distribution Lists

Subject: Call for Research Participants

Dear Staff,

I am a PhD student in the Manchester Centre for Health Economics. This is an invitation to take part in an interview that aims to understand your views about risk in healthcare and identify how balance the risks and benefits of breast cancer screening.

The interview will take approximately 45 minutes and can take place at a location convenient for you. The full information sheet about what to expect should you choose to participate is attached.

Please feel free to forward this email to anyone else, in or outside the university, who might be interested in participating in this research.

Many thanks,

Caroline Vass

PhD Student in Health Economics
The Manchester Centre for Health Economics
Institute of Population Health
University of Manchester
4.306, Jean McFarlane Building
Oxford Road
Manchester M13 9PL
Tel: +44 (0)161 306 7970
http://www.population-health.manchester.ac.uk/students/CarolineVass
Appendix 6.3: Information sheet for the eye-tracking study participants

Understanding risk in healthcare: an eye-tracking study

You are being invited to take part in a research study that aims to understand how people complete surveys. The example we will use is choosing a screening programme for breast cancer. Before you decide whether to take part in this study it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully. Please ask if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part. Thank you for reading this.

What is the aim of the research?
This study wants to understand how people make choices about healthcare. The aim is to explore people balance the risks and benefits when making a decision about breast cancer screening.

Why have I been chosen?
This study aims to explore the views of women. Any female member of the public who is able to complete a survey in English is able to take part in this study.

What would I be asked to do if I took part?
You will be invited to partake in an eye-tracking study at the University of Manchester. The study should take no more than 15 minutes. During the study I will ask you to complete a questionnaire with different imaginary breast cancer screening programmes and you will be asked which you prefer. Whilst you are completing the questionnaire, your eye-movements will be monitored, which is like having an eye-test at an opticians. At the end of the questionnaire I will ask you to complete a short survey containing questions about the choices you made.
What happens to the data collected?
The eye-tracking data will be collected to see if there are patterns in people’s answers and where they were looking whilst completing the survey. The answers to the survey will be used to help us understand how breast cancer screening programmes can be better provided. The survey questions will help us understand why different people might make different choices. Your name will not be recorded and instead you will be given an anonymous identification number.

How is confidentiality maintained?
The law called the Data Protection Act (1998) tells us how to keep your information secure. If you provide personal information (name, address, contact details) it will remain confidential and we will not give your details to anyone else.

Will I be paid for participating in the research?
We are not able to pay you for taking part in this study. We can reimburse your travel expenses for your journey to the campus in line with university policy. To thank you for your time, you will receive an Amazon voucher for participating in this research.

Where will the research be conducted?
The research will take place on The University of Manchester’s main campus, in the Zochonis Building on Brunswick Street. This building is number 60 on the campus maps which are located around the campus.

Will the outcomes of the research be published?
The main outcome of this research will be a PhD thesis for Caroline Vass, a student based in the Manchester Centre for Health Economics at The University of Manchester. In addition, we may want to report the findings at conferences or in a published journal article.

My question is not covered here, who do I contact for more information?
Please contact Caroline Vass (PhD Student in the Centre for Health Economics) on 077 1575 1262 or email caroline.vass@postgrad.manchester.ac.uk.

Thank you for reading this information sheet
Please keep this information sheet for your records
Appendix 6.4: A description of the eye-tracking pilot study

A pilot eye-tracking DCE was designed to understand the preferences of female members of the public (recruited by posters in local cafes) for a breast screening programme described by three attributes (probability of detecting a cancer, risk of unnecessary treatment, and out-of-pocket cost) each with four levels. Two survey versions were used that varied how the risk attributes (probability of detecting a cancer and risk of unnecessary treatment) were presented as: (1) a percentage or (2) a percentage and icon array. Eye-movements were recorded as a series of co-ordinates 1,000 times a second. Eye-tracking data were analysed in terms of direction of motion and total visual attention (dwell time) to pre-defined areas of interest using descriptive statistics. Immediately after completing the last choice question, respondents were asked a series of debriefing questions. The effect of each attribute on the women’s preferences were analysed using a conditional logit model.

Fifteen completed the DCE in the eye-tracking experiment. Results of the pilot study found respondents gave significantly more visual attention, indicating information processing, to both risk attributes when risk was communicated with an icon array rather than solely as a percentage with a mean dwell time of 6316 and 5043 milliseconds, respectively. Respondents to the icon array version also exhibited significantly more upwards and downwards eye-movements (43% v 38% of saccades) suggesting calculations were made in line with expected utility theory possibly reflecting a greater understanding of the risk information. The eye-tracking data confirmed the self-reported attribute non-attendance as stated by respondents when asked the de-briefing questions with significantly lower (by almost 70%) mean dwell times to these attributes. The results of the conditional logit revealed both probability of detecting a cancer and the risk of unnecessary treatment were significant in women’s decision to partake in breast screening.
Appendix 6.5: Consent form for participants in the eye-tracking study

Consent form
Understanding risk in healthcare

1. I confirm that I have read the attached information sheet on the above project and have had the opportunity to consider the information and ask questions and had these answered satisfactorily.

2. I understand that my participation in the study is voluntary and that I am free to withdraw at any time without giving a reason.

3. I agree to the use of anonymous quotes in academic articles.

I agree to take part in the above project

__________________________________________  ____________  _______________________
Name of participant                      Date                      Signature

__________________________________________  ____________  _______________________
Name of person taking consent              Date                      Signature
Appendix 6.6: iPad survey for the eye-tracking study

How confident are you that you would make the same choices if faced with the situations in real-life?

- Very confident I would make the same choices
- Quite confident I would make the same choices
- Not confident I would make the same choices
- Confident I would make different choices

Please explain your answer

Progress Bar: 0% - 100%
On a scale of 1-5, how easy or difficult did you find making choices between the alternatives?

<table>
<thead>
<tr>
<th></th>
<th>1 (Very easy)</th>
<th>2</th>
<th>3 (Neither difficult nor easy)</th>
<th>4</th>
<th>5 (Very difficult)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feedback</td>
<td>1=1</td>
<td>1=2</td>
<td>Feedback</td>
<td>1=3</td>
<td>Feedback</td>
</tr>
</tbody>
</table>

Please explain your answer

![Feedback Group]

![Feedback Group]
On a scale of 1-5, how easy or difficult did you find the survey to understand?

<table>
<thead>
<tr>
<th></th>
<th>1 (Very easy)</th>
<th>2 (Neither difficult nor easy)</th>
<th>3</th>
<th>4</th>
<th>5 (Very difficult)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feedback1</td>
<td>Feedback2</td>
<td>Feedback3</td>
<td>Feedback4</td>
<td>Feedback5</td>
</tr>
</tbody>
</table>

Was there anything in the survey that was particularly unclear?

No, I understood everything.

There were some things that weren't clear or well explained.

(please comment in the box below)

Is there anything else about the survey we could change to make it easier to understand (e.g. phrasing, explanations)?
Did you find yourself making choices based on one or two characteristics of the screening programme?

Yes, I based choices on one or two characteristics

No, I considered ALL characteristics

Would you say you looked at the characteristic women who have a cancer detected?

Please select

Would you say you looked at the characteristic risk of unnecessary treatment?

Please select

Would you say you looked at the characteristic cost?

Please select

If you based your choices on one or two characteristics, why was this? (Please tick all that apply):

There was just too many characteristics on which to make a decision

The other characteristic(s) were too complicated or it was unclear what it meant

Other reason (please explain)
What is your occupational status?

- Employed full-time
- Employed part-time
- Self-employed
- Unemployed
- Retired
- Looking after a home/family
- Student
- Freelance or temping
- Long-term sickness
- Temporarily laid off

What is your occupation?
(Or previous occupation if retired?)

---
What is the highest level of education you have obtained?

- No formal qualifications
- 1-4 O-levels/GCSEs
- 5+ O-levels/GCSEs
- NVQs
- A-levels/AS-levels
- Undergraduate degree
- Master’s degree
- PhD
- Other formal qualification
What is your ethnic group?

- White British/Irish
- White other
- Mixed
- Black/Black British
- Asian/Asian British
- Other (please specify)
What is your religious group?

- No religion
- Christian
- Buddhist
- Jewish
- Hindu
- Muslim
- Sikh
- Other (please specify)
Do you have any children?

Yes

No
Do you have any daughters?

- [ ] Yes
- [ ] No
Have you ever had breast screening (a mammogram)?

- ExperienceA=1: Yes
- ExperienceA=2: No
- ExperienceA=3: Don't know

Have you received follow-up (such as an ultrasound, a repeat mammogram or a biopsy) for breast cancer?

- ExperienceB=1: Yes
- ExperienceB=2: No
- ExperienceB=3: Don't know
Thinking about the last time you went for breast screening, what was your main form of transport to the screening?

Please select as appropriate.

- walked
- cycled
- travelled by bus
- travelled by train/metro
- travelled by private car
- travelled by taxi
- travelled by motorbike
- other transport [ ]

Still thinking about your last breast screen, how much did it cost you for your return journey (even if your start and end points were different)?
(If you did not pay any transport fare then please enter 0)

[ ]

Did you incur any other “costs”?

- I took time off work
- I had to pay for care (for children, relative, friend or other dependent)
- I lost income from my own business
- other [ ]
Have you, or has anyone in your family (blood relative or otherwise) had breast cancer?

- ExperienceC
  - ExperienceC=1 Yes
  - ExperienceC=2 No
  - ExperienceC=3 Don't know

Do you know of anybody else (e.g. friends or colleagues) who has had breast cancer?

- ExperienceC1=1 Yes
- ExperienceC1=2 No
- ExperienceC1=3 Don't know

0% 100%
How would you describe your current risk of breast cancer?

- High or very high
- Quite high/above average
- Average
- Slight or low
- No idea

How would you describe your future risk of breast cancer?

- High or very high
- Quite high/above average
- Average
- Slight or low
- No idea

To what extent are you concerned about your own risk of breast cancer?

- Not at all
- A little
- Quite a lot
- Very much
- No idea

Is there anything else you would like to tell us about your risk of breast cancer?

[Blank field]
On a scale of 1-5, how much do you agree with the following statements?

"I feel life is very much out of our hands, and fate determines everything"

<table>
<thead>
<tr>
<th></th>
<th>1 (Strongly disagree)</th>
<th>2 (Disagree)</th>
<th>3 (Neither agree nor disagree)</th>
<th>4 (Agree)</th>
<th>5 (Strongly agree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FeedbackA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeedbackB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeedbackC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

"I get a thrill when I gamble, even if it's just the lottery"

<table>
<thead>
<tr>
<th></th>
<th>1 (Strongly disagree)</th>
<th>2 (Disagree)</th>
<th>3 (Neither agree nor disagree)</th>
<th>4 (Agree)</th>
<th>5 (Strongly agree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FeedbackA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeedbackB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeedbackC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

"Even if something is unlikely, I feel it will probably happen to me"

<table>
<thead>
<tr>
<th></th>
<th>1 (Strongly disagree)</th>
<th>2 (Disagree)</th>
<th>3 (Neither agree nor disagree)</th>
<th>4 (Agree)</th>
<th>5 (Strongly agree)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FeedbackA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeedbackB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FeedbackC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Percentage: 100%**
This next section will ask you some questions to understand how familiar you are with probabilities.

If you are unsure of the answer, you can skip the question.

Imagine I flip a coin 1,000 times. What is your best guess about how many times the coin would fall heads up in 1,000 flips? 

If there was a lottery where the chance of winning £10 is 1%, if 1,000 people bought a ticket, how many people would you expect to win? 

In another lottery, the chance of winning a car is 1 in 1,000. What percentage of tickets in the lottery will win a car?
You missed some of the previous questions about statistics and probability. Why was this?

- Just too hard
- We're going to take me too long to answer
- Can't see the relevance to the study
- Other

0% 100%
Can you think of anything else which might help us understand how you made your choices in the choice task?
Thank you for completing this survey.

If you have been affected by any of the issues raised in this survey, the following organisations may be of assistance:

- National Hereditary Breast Cancer Helpline on 01629 813000
- Breast Cancer Care on 0808 800 6000

If you have any questions or concerns please contact Caroline Vasa on 077 1675 1262 or caroline.vasa@postgrad.manchester.ac.uk.

Manchester Centre for Health Economics, Jean McFarlane Building, University of Manchester, Oxford Road, Manchester M13 9PL
Appendix 6.7: Heteroskedastic conditional logit results for the eye-tracking study

<table>
<thead>
<tr>
<th>Heteroskedastic conditional logit model (linear risk, pooled sample)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Utility</strong></td>
<td></td>
</tr>
<tr>
<td>ASC (none)</td>
<td>-1.565*** (0.41)</td>
</tr>
<tr>
<td>Detect</td>
<td>0.141*** (0.03)</td>
</tr>
<tr>
<td>Risk</td>
<td>-0.076*** (0.02)</td>
</tr>
<tr>
<td>Cost</td>
<td>-0.031 (0.03)</td>
</tr>
<tr>
<td>IAP*detect</td>
<td>1.789 (4.29)</td>
</tr>
<tr>
<td>IAP*risk</td>
<td>-0.901 (2.07)</td>
</tr>
<tr>
<td><strong>Scale</strong></td>
<td></td>
</tr>
<tr>
<td>Icon arrays and percentages(IAP)</td>
<td>-2.152 (2.12)</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>n</td>
</tr>
<tr>
<td>-329.87321</td>
<td>42</td>
</tr>
<tr>
<td>Obs</td>
<td>1386</td>
</tr>
</tbody>
</table>

* p<0.005; ** p<0.01; ***p<0.001; standard errors in parentheses