

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/172039>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

**Derivation of preliminary preference-based valuation
sets for the Short Warwick-Edinburgh Mental
Wellbeing Scale (SWEMWBS) to allow calculation of
Mental Well-being Adjusted Life Years (MWALY)**

by

Hei Hang Edmund Yiu

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Health Sciences

Centre for Health Economics at Warwick

Warwick Medical School

University of Warwick

March 2022

Contents

| | |
|---|----|
| Acknowledgements..... | 11 |
| Declaration..... | 12 |
| Abstract..... | 13 |
| List of Abbreviations | 14 |
| Chapter 1: Introduction..... | 18 |
| 1.1. Background..... | 18 |
| 1.1.1. Concepts of well-being | 19 |
| 1.1.1.1. Hedonic well-being..... | 19 |
| 1.1.1.2. Eudaimonic well-being..... | 19 |
| 1.1.1.3. Objective list theory..... | 20 |
| 1.1.1.4. Desire fulfilment (or preference satisfaction) theory..... | 20 |
| 1.1.1.5. Economic well-being..... | 20 |
| 1.1.1.6. Capability theory..... | 21 |
| 1.2. Objective and research questions..... | 22 |
| 1.3. Structure of thesis..... | 22 |
| Chapter 2: Review of measurement instruments | 24 |
| 2.1. Introduction..... | 24 |
| 2.2. Theoretical concepts of welfarism and extra-welfarism | 24 |
| 2.2.1. Welfarism..... | 25 |
| 2.2.1.1. Application in economic evaluation | 26 |
| 2.2.2. Extra-welfarism..... | 27 |
| 2.2.2.1. Application in economic evaluation | 27 |
| 2.2.3. The capability approach..... | 28 |
| 2.2.4. Summary | 29 |
| 2.3. A comparative analysis of existing generic preference-based measurement instruments adopted across the world..... | 29 |
| 2.3.1. Description of preference-based instruments..... | 30 |
| 2.3.1.1. MAU measures | 30 |
| 2.3.1.2. Preference-based capability measures | 31 |
| 2.3.2. Concepts and constructs covered within the questions of the identified preference-based instruments..... | 31 |
| 2.3.2.1. Comments on the coverage of physical health dimensions | 34 |
| 2.3.2.2. Comments on the coverage of mental health dimensions..... | 34 |
| 2.3.2.2.1. MAU instruments - Constructs and their applications or coverage of mental health dimensions..... | 34 |
| 2.3.2.2.1.1. QWB-SA..... | 34 |

| | | |
|--------------|--|----|
| 2.3.2.2.1.2. | EQ-5D-5L | 35 |
| 2.3.2.2.1.3. | HUI3 | 36 |
| 2.3.2.2.1.4. | SF-6D..... | 37 |
| 2.3.2.2.1.5. | 15D..... | 39 |
| 2.3.2.2.1.6. | AQoL-8D | 40 |
| 2.3.2.2.1.7. | ReQoL..... | 41 |
| 2.3.2.2.2. | Preference-based capability instruments - Constructs of mental health dimensions | 42 |
| 2.3.2.2.2.1. | ICECAP-A and ICECAP-O..... | 42 |
| 2.3.2.2.2.2. | ASCOT | 43 |
| 2.3.2.3. | Overview of instruments with regard to measuring mental well-being..... | 43 |
| 2.4. | A comparative analysis of non-preference-based mental well-being measures..... | 45 |
| 2.4.1. | WEMWBS/SWEMWBS | 45 |
| 2.4.2. | WHO-5..... | 46 |
| 2.4.3. | MHC-SF..... | 46 |
| 2.4.4. | SEHS..... | 47 |
| 2.4.5. | Summary | 47 |
| 2.5. | Justification for the need to estimate the MWALY through the development of a preference-based tariff for a mental well-being instrument..... | 48 |
| 2.6. | Discussion of the best choice of instrument for preference elicitation | 50 |
| 2.6.1. | Evidence to support the use of WEMWBS and/or SWEMEBS in measuring mental well-being..... | 52 |
| 2.7. | Conclusion | 52 |
| Chapter 3: | An overview of research methods for constructing the valuation set | 53 |
| 3.1. | Introduction..... | 53 |
| 3.2. | The six stages of the construction of a preference-based tariff for the WEMWBS/SWEMWBS | 53 |
| 3.2.1. | Stage I: Establish dimensions | 54 |
| 3.2.2. | Stage II: Eliminate and select the best items per dimension | 54 |
| 3.2.3. | Stage III: Explore item-level reduction..... | 56 |
| 3.2.4. | Stage IV: Validation: repeat stages I to III on other data sets..... | 56 |
| 3.2.5. | Stages V and VI: Valuation exercise to elicit state values for a sample of states & model valuation results to produce utility values for all states | 57 |
| 3.2.5.1. | Identification of the appropriate valuation techniques for mental well-being states | 57 |
| 3.2.5.1.1. | Direct valuation methods | 57 |
| 3.2.5.1.2. | Indirect valuation methods..... | 58 |
| 3.2.5.1.3. | Justification of the valuation techniques for SWEMWBS..... | 58 |

| | | |
|----------------|--|----|
| 3.2.5.1.3.1. | Mapping | 58 |
| 3.2.5.1.3.2. | Visual analogue scale (VAS) | 66 |
| 3.2.5.1.3.3. | Magnitude estimation (ME) | 66 |
| 3.2.5.1.3.4. | Person trade-off (PTO) | 66 |
| 3.2.5.1.3.5. | Standard gamble (SG) | 67 |
| 3.2.5.1.3.6. | Time trade-off (TTO) | 67 |
| 3.2.5.1.3.6.1. | Lead-time versus lag-time TTO | 68 |
| 3.2.5.1.3.6.2. | Time horizon and duration of well-being state | 70 |
| 3.2.5.1.3.6.3. | Iteration algorithm | 73 |
| 3.2.5.1.3.7. | Discrete choice experiments (DCEs) and Best-worst scaling (BWS) | 73 |
| 3.2.5.1.4. | Administrative technology for the valuation procedure | 75 |
| 3.2.5.2. | Selection of a sample of mental well-being states for valuation | 78 |
| 3.2.5.2.1. | Design for the DCE | 79 |
| 3.2.5.2.2. | Design for the C-TTO | 79 |
| 3.2.5.3. | Piloting studies to validate the valuation methodology in a suitable sample | 80 |
| 3.2.5.3.1. | Phase I (Qualitative phase): Cognitive interviews with the use of think- aloud and verbal probing techniques | 81 |
| 3.2.5.3.2. | Phase II (Quantitative phase): Structured interviews to test the psychometric or empirical properties of valuation protocol | 83 |
| Chapter 4: | Cognitive interviews for the qualitative validation of valuation protocol | 84 |
| 4.1. | Introduction | 84 |
| 4.2. | Methods | 84 |
| 4.2.1. | Recruitment of respondents | 86 |
| 4.2.2. | Sample size | 87 |
| 4.2.3. | Experimental design for the selection of SWEMWBS states | 87 |
| 4.2.3.1. | Design for the DCE | 87 |
| 4.2.3.2. | Design for the C-TTO | 88 |
| 4.2.4. | Valuation platform | 90 |
| 4.2.5. | Interview process | 90 |
| 4.2.6. | Data analysis | 94 |
| 4.3. | Results | 95 |
| 4.3.1. | Theme 1: Format and structure | 96 |
| 4.3.1.1. | Inappropriate examples | 96 |
| 4.3.1.2. | Increase in the variety of preliminary assessments | 97 |
| 4.3.1.3. | Confusion on scenario completion | 97 |
| 4.3.1.3.1. | Mistakenly clicking the non-preferred life | 97 |

| | | |
|------------|--|-----|
| 4.3.1.3.2. | Failure to adjust time properly | 97 |
| 4.3.1.3.3. | Clarification of meaning of a state..... | 98 |
| 4.3.1.3.4. | System operation issue..... | 98 |
| 4.3.1.4. | Improvement of presentation layout | 98 |
| 4.3.1.4.1. | C-TTO Feedback Module | 98 |
| 4.3.1.4.2. | Flow of the interview | 99 |
| 4.3.2. | Theme 2: Items and levels | 99 |
| 4.3.2.1. | Contradiction in levels | 99 |
| 4.3.2.2. | Compensation effect | 100 |
| 4.3.2.3. | Overlapping effect | 100 |
| 4.3.2.4. | Non-linear effects of levels..... | 101 |
| 4.3.2.5. | Inferiority of top levels | 101 |
| 4.3.3. | Theme 3: Decision strategies | 101 |
| 4.3.3.1. | Lexicographic ordering..... | 101 |
| 4.3.3.2. | Interpretation of levels | 102 |
| 4.3.3.3. | Comparison with previous tasks | 102 |
| 4.3.3.4. | Personal and external factors | 103 |
| 4.3.3.5. | Availability heuristic..... | 105 |
| 4.3.3.6. | Duration of C-TTO states | 106 |
| 4.3.3.7. | Satisficing heuristic..... | 106 |
| 4.3.3.8. | Ignorance of identical levels of attributes between DCE alternatives | 106 |
| 4.3.3.9. | Rejection of unimaginable states | 107 |
| 4.3.3.10. | Framing effect..... | 107 |
| 4.3.3.11. | Integration of self-written notes..... | 108 |
| 4.3.4. | Theme 4: Valuation feasibility..... | 108 |
| 4.3.5. | Theme 5: Valuation outcome..... | 110 |
| 4.3.5.1. | Failure to reach the C-TTO indifference point | 110 |
| 4.3.5.2. | Non-trading effects | 111 |
| 4.3.6. | Theme 6: Overall Reflections on mental well-being | 111 |
| 4.4. | Discussion | 112 |
| 4.4.1. | Format and structure | 113 |
| 4.4.2. | Items and levels..... | 115 |
| 4.4.3. | Decision strategies | 116 |
| 4.4.4. | Valuation feasibility..... | 118 |
| 4.4.5. | Valuation outcome | 120 |
| 4.4.6. | Overall reflections on mental well-being..... | 120 |
| 4.5. | Conclusion | 121 |

| | |
|---|-----|
| Chapter 5: A quantitative investigation of the feasibility, practicality and face validity of the C-TTO and DCE in the valuation of SWEMWBS | 122 |
| 5.1. Introduction..... | 122 |
| 5.2. Methods..... | 122 |
| 5.2.1. Recruitment strategy | 122 |
| 5.2.2. Experimental design and sample size determination | 123 |
| 5.2.2.1. Design for the DCE..... | 123 |
| 5.2.2.2. Design for the C-TTO..... | 123 |
| 5.2.3. Analysis..... | 125 |
| 5.2.3.1. Feasibility and practicality..... | 125 |
| 5.2.3.2. Face validity..... | 127 |
| 5.2.4. Interview process | 127 |
| 5.3. Results..... | 128 |
| 5.3.1. Feasibility and practicality | 132 |
| 5.3.1.1. C-TTO..... | 132 |
| 5.3.1.2. DCE | 138 |
| 5.3.1.3. Overall impression..... | 141 |
| 5.3.2. Face validity..... | 142 |
| 5.3.2.1. C-TTO..... | 142 |
| 5.3.2.2. DCE | 144 |
| 5.4. Discussion..... | 146 |
| 5.5. Conclusion | 152 |
| Chapter 6: Modelling preliminary versions of preference-based valuation set | 153 |
| 6.1. Introduction..... | 153 |
| 6.2. Methods..... | 153 |
| 6.2.1. Heteroskedastic Tobit model for the C-TTO data..... | 153 |
| 6.2.2. Conditional Logit model for the DCE data | 155 |
| 6.2.2.1. Rescaling..... | 156 |
| 6.2.2.1.1. Anchoring to the lowest mental well-being state of the C-TTO | 156 |
| 6.2.2.1.2. Mapping DCE onto C-TTO | 157 |
| 6.2.2.1.3. Hybrid model (the EuroQol hybrid model)..... | 158 |
| 6.2.3. Inverse Variance Weighting (IVW) approach for both the C-TTO and DCE data | 158 |
| 6.2.4. Description of explanatory variables | 160 |
| 6.2.5. Model analysis | 162 |
| 6.2.6. Sensitivity analysis..... | 163 |
| 6.3. Results..... | 163 |

| | | |
|-----------------|---|-----|
| 6.3.1. | The C-TTO models | 163 |
| 6.3.2. | The DCE models..... | 170 |
| 6.3.2.1. | Anchoring to the lowest mental well-being state of the C-TTO..... | 174 |
| 6.3.2.2. | Mapping DCE onto C-TTO | 175 |
| 6.3.2.3. | The EuroQol hybrid model | 179 |
| 6.3.3. | The IVW hybrid model | 179 |
| 6.3.4. | Sensitivity analysis..... | 183 |
| 6.3.5. | Comparison of valuation sets..... | 184 |
| 6.4. | Discussion | 193 |
| 6.5. | Conclusion | 197 |
| Chapter 7: | Discussion and conclusion | 198 |
| 7.1. | Introduction..... | 198 |
| 7.2. | Summary and discussion of main results to the research questions..... | 198 |
| 7.2.1. | Do any existing preference-based measurement approaches and instruments value mental well-being? | 198 |
| 7.2.2. | Are there any mental well-being measures that can be used to develop a preference-based tariff?..... | 199 |
| 7.2.3. | What is the best choice of instrument for the elicitation of a preference-based tariff to allow the calculation of MWALYs? | 199 |
| 7.2.4. | What is the appropriate valuation protocol for the valuation of mental well-being state?..... | 200 |
| 7.3. | Application and role of the valuation sets..... | 206 |
| 7.4. | Contributions of this research | 213 |
| 7.5. | Limitations and directions for future research | 213 |
| 7.6. | Conclusion | 219 |
| Appendices..... | | 220 |
| Appendix 1: | Description of the MAU instruments..... | 220 |
| Appendix 2: | Description of the preference-based capability instruments | 226 |
| Appendix 3: | Description of the mental well-being instruments | 227 |
| Appendix 4: | Studies focusing solely on the validation of WEMWBS | 229 |
| Appendix 5: | Evidence focused solely on the validation of SWEMWBS | 233 |
| Appendix 6: | Evidence focused on the validation of both WEMWBS and SWEMWBS | 237 |
| Appendix 7: | A review of the findings covering stage II to stage IV for the development process of a mental well-being preference-based instrument | 240 |
| Appendix 8: | Descriptive system of the SWEMWBS | 243 |
| Appendix 9: | A review of the direct valuation techniques..... | 245 |
| Appendix 10: | A review of the indirect valuation techniques..... | 250 |
| Appendix 11: | An example of the advertisement layout in the qualitative phase..... | 252 |

| | |
|--|-----|
| Appendix 12: Syntax for the DCE experimental design in Ngene | 253 |
| Appendix 13: The 32 pairs of MWB states included in the DCE valuation tasks | 254 |
| Appendix 14: Code for the C-TTO experimental design in R..... | 256 |
| Appendix 15: The 50 SWEMWBS MWB states included in the C-TTO valuation tasks..... | 260 |
| Appendix 16: An additional version of C-TTO practice example | 262 |
| Appendix 17: An algorithm to explore potential highly uncommon reported SWEMWBS states | 263 |
| Appendix 17.1: The frequency and proportion of the top 10 responses | 264 |
| Appendix 17.2: States with D values between 10 and 12 | 265 |
| Appendix 17.3: The D values of implausible states claimed by participants | 269 |
| Appendix 18: An example of the advertisement for interview recruitment..... | 275 |
| Appendix 19: Syntax for the DCE experimental design in Ngene | 276 |
| Appendix 20: The 50 pairs of mental well-being states included in the DCE valuation tasks .. | 277 |
| Appendix 21: Code for the C-TTO experimental design in R..... | 280 |
| Appendix 22: The 64 SWEMWBS mental well-being states included in the C-TTO valuation tasks | 283 |
| Appendix 23: Facebook advertising statistics..... | 285 |
| Appendix 24: The EuroQol hybrid model | 288 |
| Appendix 24.1: A graphical relationship between DCE unscaled values and C-TTO utility values | 288 |
| Appendix 24.2: Modelling result of the EuroQol hybrid model..... | 288 |
| Appendix 24.3: A graphical relationship between selected C-TTO utility values and the EuroQol hybrid utility values, ordered by the C-TTO utility values..... | 291 |
| Appendix 25: Sensitivity analysis of the C-TTO data | 293 |
| Appendix 26: Sensitivity analysis of the DCE data..... | 296 |
| References..... | 299 |

List of Tables

| | |
|---|-----|
| Table 1: Comparison of the concepts of welfarism and extra-welfarism | 25 |
| Table 2: An overview of the concepts covered by the MAU and preference-based capability instruments..... | 32 |
| Table 3: A summary of the valuation strategies adopted by the MAU instruments and capability measures..... | 60 |
| Table 4: An example of a pairwise DCE with forced choice..... | 75 |
| Table 5: Summary of the valuation strategy for the SWEMWBS | 77 |
| Table 6: A summary of the C-TTO and DCE experimental designs in the qualitative phase | 89 |
| Table 7: Examples of probing questions during the think-aloud process for the C-TTO tasks... | 92 |
| Table 8: Follow-up debriefing questions if they were not addressed within the think-aloud process of the C-TTO tasks..... | 93 |
| Table 9: Examples of probing questions during the think-aloud process for the DCE tasks..... | 93 |
| Table 10: Follow-up debriefing questions if they were not addressed within the think-aloud process of the DCE tasks. | 93 |
| Table 11: Overall debriefing questions for both parts of the interview | 93 |
| Table 12: Demographic characteristics of 14 participants..... | 95 |
| Table 13: Quotes related to the influence of personal factors on preferences | 103 |
| Table 14: Issues identified by the interview and the corresponding proposed modification to the valuation protocol | 112 |
| Table 15: A summary of the C-TTO and DCE experimental designs in the quantitative phase | 124 |
| Table 16: C-TTO and DCE debriefing questions | 125 |
| Table 17: Demographic characteristics of the 225 participants..... | 129 |
| Table 18: Response statistics of the C-TTO debriefing statements | 131 |
| Table 19: Reasons for not using the remote-control function..... | 133 |
| Table 20: Failure to reach the indifference point for the worse-than-death scenario..... | 134 |
| Table 21: Failure to reach the indifference point for the better-than-death scenario..... | 135 |
| Table 22: Response statistics of the DCE debriefing statements..... | 137 |
| Table 23: Response statistics of the overall debriefing statements..... | 140 |
| Table 24: A description of the explanatory variables | 160 |
| Table 25: Results from different specifications of the C-TTO data..... | 163 |
| Table 26: Wald tests for model comparison | 170 |
| Table 27: The total weight for each attribute in the Model 1A..... | 170 |
| Table 28: Results from different specifications of the DCE data | 170 |
| Table 29: The total weight for each attribute in the Model 2A based on anchoring..... | 175 |

| | |
|---|-----|
| Table 30: Mapping result..... | 176 |
| Table 31: Calculation of utility change for the levels of attribute makeupownmind..... | 176 |
| Table 32: Utility change from level 5 for all attribute levels..... | 177 |
| Table 33: The total weight for each attribute in the Model 2A based on mapping..... | 178 |
| Table 34: Result of the IVW hybrid model..... | 179 |
| Table 35: The total weight for each attribute in Model 3 | 182 |
| Table 36: Ranking of attributes and the largest and smallest level change from level 5 | 187 |
| Table 37: Examples of utility calculation for particular states | 189 |
| Table 38: MAD and RMSD performance of the DCE rescaling methods..... | 192 |
| Table 39: The efficacy of applying the modified valuation protocol to the quantitative phase. | 202 |
| Table 40: The influence of tariff choice on the effectiveness of interventions with identical implementation costs and change in outcome score | 207 |
| Table 41: Sub-group analysis for the effectiveness of an intervention with different magnitude changes in outcome score between groups | 210 |
| Table 42: Sub-group analysis for the effectiveness of an intervention with identical magnitude change in level score for the same attribute between groups..... | 211 |

List of Figures

| | |
|--|-----|
| Figure 1: The six stages of the development of a preference-based instrument | 53 |
| Figure 2: The C-TTO layout | 71 |
| Figure 3: The C-TTO task..... | 85 |
| Figure 4: A pairwise DCE with forced choice | 86 |
| Figure 5: A visual presentation of the C-TTO Feedback Module..... | 92 |
| Figure 6: An example of a DCE pair with identical levels for two items | 107 |
| Figure 7: Distribution of the C-TTO values..... | 143 |
| Figure 8: Relationship between mean C-TTO value and level-sum score..... | 144 |
| Figure 9: Relationship between the percentage of the chosen option and the difference in level-sum score among the 2246 responses | 145 |
| Figure 10: Relationship between the standard deviation of the C-TTO responses against the level-sum score | 155 |
| Figure 11: Relationship between the mean C-TTO values and the unscaled DCE values for the 64 states valued by the participants..... | 158 |
| Figure 12: The utility values generated by the C-TTO model, two DCE rescaling models, and the IVW hybrid model, against the SWEMWBS states ordered by the C-TTO utility values | 190 |

Acknowledgements

Firstly, I would like to express sincere thankfulness for my three supervisors: Professor Jason Madan, Professor Stavros Petrou, and Professor Sarah Stewart-Brown. They provided continuous support throughout my PhD journey in terms of funding, research training, and mental health management. Also, thank you Professor Hareth Al-Janabi from the University of Birmingham for providing experienced advice on coding and thematic analysis in my qualitative phase. Thank you Dr. John Buckell from the University of Oxford for providing professional advice on experimental designs and choice modelling analyses in my quantitative phase. I won't be able to reach this stage of producing high-quality research outputs without the constructive feedback and guidance from all of them.

I would like to acknowledge attendees at the Health Economists' Study Group Summer (HESG) 2021 Meeting and the 2021 International Health Economics Association (iHEA) Congress for discussing and providing insights on the interpretation of result in my qualitative phase. I was also delighted to have the EQ-PVT platform shared by the EuroQol Group for recording the valuation responses in both my qualitative and quantitative phase.

Lastly, I was grateful to receive Health Economics teaching opportunities at the Economics Department, referred by Professor Jacob Glazer. He provided trust and recognition on my teaching potential, boosting my confidence and satisfaction level on teaching.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Professor Hareth Al-Janabi was the second rater and Professor Jason Madan was the third rater for checking the coding of a few transcripts in the qualitative phase (Chapter 4).

Part of the contents in the qualitative phase (Chapter 4) was submitted to the Health Economists' Study Group Summer 2021 Meeting as a conference paper.

Part of the contents in the qualitative phase (Chapter 4) has been published in Quality of Life Research.

Part of the contents in this thesis was submitted as an upgrade report during my first year.

Abstract

Concerns have been raised regarding the sensitivity of widely used preference-based instruments (e.g. EQ-5D) to value mental health benefits. An alternative outcome measure other than QALY is required due to an increasing interest in the promotion of mental well-being.

The aim of this thesis is to develop preliminary U.K. preference-based valuation sets for the Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS), to allow estimation of Mental Well-being Adjusted Life Years (MWALYs).

Given that this was the first attempt at valuing SWEMWBS states, a series of stages were followed to ensure the robustness of the derived valuation sets. Firstly, alternative valuation techniques were analysed to justify the appropriate valuation strategy for mental well-being states. A sample of manageable mental well-being states for valuation was also identified through alternative experimental designs. Next, a qualitative piloting study with the application of think-aloud interviewing technique was conducted to investigate the cognitive process of completing the valuation tasks. The modified valuation protocol informed by the qualitative study was then validated within a larger sample in a quantitative study. The valuation responses of the quantitative study were modelled to produce utility values for all mental well-being states.

The qualitative and quantitative studies suggested the feasibility, practicality and face validity of the SWEMWBS valuation. A total of 225 participants provided valuation responses to allow estimation of valuation sets based on composite time trade-off (C-TTO), Discrete choice experiment (DCE) and Inverse Variance Weighting (IVW) hybrid model. The first and second models generated illuminating differences with the hybrid approach giving an arguably desirable blend of the two. The valuation sets for mental well-being states can be used for indicative cost-utility analyses of mental well-being interventions and have the potential to inform the practicality of applying the proposed valuation protocol in the full national valuation study.

List of Abbreviations

| <u>Abbreviation</u> | <u>Explanation</u> |
|---------------------|---|
| AQoL | Assessment of Quality of Life measure |
| AQoL-4D | Assessment of Quality of Life - 4D |
| AQoL-6D | Assessment of Quality of Life - 6D |
| AQoL-7D | Assessment of Quality of Life - 7D |
| AQoL-8D | Assessment of Quality of Life - 8D |
| ASCOT | The Adult Social Care Outcomes Toolkit |
| BDI | Beck Depression Inventory |
| BWS | Best-worst scaling |
| BSL | British Sign Language |
| CAPI | Computer-Assisted Personal Interview |
| CFI | Comparative Fit Index |
| CHEW | Centre for Health Economics at Warwick |
| CI | Confidence interval |
| CIDI | Composite International Diagnostic Interview |
| CORE-OM | Clinical Outcomes in Routine Evaluation - Outcome Measure |
| CORE-6D | Clinical Outcomes in Routine Evaluation - 6D |
| C-SWEMWBS | Chinese version of Short Warwick-Edinburgh Mental Wellbeing Scale |
| C-TTO | Composite time trade-off |
| 15D | The 15-dimensional instrument |
| DCE | Discrete choice experiment |

| | |
|----------|--|
| EQ-5D | EuroQol five dimension measure |
| EQ-5D-3L | EuroQol five dimension 3-level version |
| EQ-5D-5L | EuroQol five dimension 5-level version |
| EQHWB-S | EuroQol Health and Wellbeing Short |
| EQ-PVT | EuroQol Portable Valuation Technology |
| EQ-VT | EuroQol Valuation Technology |
| GAD-7 | General Anxiety Disorder – 7 |
| GDP | Gross Domestic Product |
| GHQ-12 | General Health Questionnaire -12 |
| HRQoL | Health-related quality of life |
| HRSD-17 | Hamilton Rating Scale for Depression |
| HUI | Health Utilities Index |
| HUI1 | Health Utilities Index Mark 1 |
| HUI2 | Health Utilities Index Mark 2 |
| HUI3 | Health Utilities Index Mark 3 |
| ICECAP-A | Investigating Choice Experiments Capability Measure for Adults |
| ICECAP-O | Investigating Choice Experiments Capability Measure for Older people |
| IVW | Inverse Variance Weighting |
| K6 | Non-specific psychological distress |
| MAU | Multi-attribute utility |
| ME | Magnitude estimation |
| MHC-SF | Mental Health Continuum-Short Form |

| | |
|----------|---|
| MVH | Measurement and Valuation of Health |
| MWALY | Mental Well-being Adjusted Life Year |
| NICE | National Institute for Health and Care Excellence |
| PHQ-9 | Patient Health Questionnaire - 9 |
| PTO | Person trade-off |
| QALY | Quality-Adjusted Life Year |
| QoL | Quality of life |
| QWB | Quality of Well-Being scale |
| QWB-SA | The Quality of Well-Being Scale Self-Administered |
| ReQoL | Recovering Quality of Life |
| ReQoL-UI | Recovering Quality of Life - Utility Index |
| ReQoL-10 | Recovering Quality of Life - 10 |
| ReQoL-20 | Recovering Quality of Life - 20 |
| RMSEA | Root mean square error of approximation |
| SCL-90 | Symptom Checklist |
| SD | Standard deviation |
| SDQ | Strengths and Difficulties Questionnaire |
| SEHS | Social and Emotional Health Survey |
| SEM | Standard error of measurement |
| SF-6D | Short-Form Six-Dimension |
| SF-36 | Short-Form 36 Health Survey |
| SG | Standard gamble |
| SQLS | Schizophrenia Quality of Life Scale |
| SWEMWBS | Short Warwick-Edinburgh Mental Wellbeing Scale |

| | |
|--------|--|
| TLI | Tucker-Lewis index |
| TTO | Time trade-off |
| VAS | Visual analogue scale |
| WEMWBS | Warwick-Edinburgh Mental Wellbeing Scales |
| WHO-5 | The World Health Organisation- Five Well-Being Index |
| WMS | Warwick Medical School |

Chapter 1: Introduction

1.1. Background

Economic evaluation is defined as the comparative analysis of alternative courses of action in terms of both their costs and consequences (Michael F. Drummond *et al.*, 2015). It can be used to inform decision making in the allocation of resources within the healthcare sector, and across the public sector more broadly, and to identify the value for money of competing interventions (I Aniza, 2008). Given that publicly funded resources are scarce and limited in supply relative to unlimited human wants, decision makers face trade-offs and associated opportunity costs (foregone benefits) in their decisions as the allocation of resources in a certain area will reduce the availability of resources for other areas. Because of this, discussions about effective and efficient ways of utilising finite resources have arisen as the scarcity of resources is a constant factor underpinning healthcare decision-making processes. With a view to generating accurate estimates of the cost-effectiveness and economic value of an intervention, an appropriate generic outcome measure in economic evaluation is required to reflect the benefits of publicly funded healthcare interventions. Also, more broadly speaking in terms of public health which aims to promote health and prevent diseases through the activities of a wide range of sectors beyond healthcare sector, an outcome measure which could capture public healthcare value is required.

The quality-adjusted life year (QALY) is a preference-based outcome measure that combines health-related quality of life (HRQoL) and length of life in a single metric and has been used to inform the cost-effectiveness of competing healthcare interventions, acting as a key data input into the incremental cost-effectiveness ratio that informs healthcare decision-making. However, the instruments widely used to derive the HRQoL component of the QALY are subject to limitations as they measure and value benefits of interventions related to certain “health” dimensions, without sufficiently capturing other aspects of well-being that may be relevant to individuals and decision-makers.

1.1.1. Concepts of well-being

Well-being can be broadly interpreted as “how well a life is going”. However, it is more than the absence of disease (being ill). There are generally two traditional classifications of well-being: hedonic and eudaimonic views.

1.1.1.1. Hedonic well-being

The hedonic view of well-being is based on the idea that pleasure and pain are the two elements determining the level of well-being (Hooker, 2015; McMahan & Estes, 2011; Ryan & Deci, 2001). Intrinsically, pleasure is a good and pain is a bad. An individual can enjoy a higher level of well-being under the experience of increased pleasure and/or decreased pain. The pursuit of optimal well-being is achieved through the greatest balance of pleasure over pain (Kahneman *et al.*, 1999; Parducci, 1995). As hedonism advocates the importance of *feelings* and *moods* for the assessment of well-being, it is usually termed subjective well-being, which is about the evaluations of an individual’s life in terms of cognitive aspects (life satisfaction) and affective aspects (emotion) (Diener, 1984; Heginbotham & Newbigging, 2013).

1.1.1.2. Eudaimonic well-being

Due to increasing concern that hedonism is insufficient to reflect all aspects of well-being when it is based solely on happiness or life satisfaction, an alternative approach named the eudemonic view of well-being was firstly proposed by Aristotle, an ancient Greek philosopher (Adler & Seligman, 2016; McMahan & Estes, 2011; Ryan & Deci, 2001). Within this view of well-being, true happiness is attained by living a virtuous life and doing what is worth doing. It is concerned with different dimensions of positive psychological functionings, including positive relations with the others, autonomy, etc (Ryff, 1995). Self-actualization is the ultimate goal of a human life in which an individual should engage in meaningful activities in order to realise the meaning of life and identify the true self. It is worth noting that most authorities now regard well-being as covering both hedonic and eudemonic approaches (Compton *et al.*, 1996; King & Napa, 1998; Ryan & Deci, 2001).

In addition to the two broad distinct views of well-being, a myriad of well-being theories have been developed to investigate different elements and schools of well-being. Some of the mainstream theories are mentioned below:

1.1.1.3. Objective list theory

In addition to the focus on the fulfilment of desires and the maximisation of pleasure, this theory strives for the identification of a list of goods that contribute benefits to an individual (Brey, 2012; Hooker, 2015). The level of well-being increases when an individual possesses more prudential goods within the list, regardless of personal attitudes or tastes (Rice, 2013). Several lists have been developed by scholars including Derek Parfit, John Finnis, Mark Murphy and Guy Fletcher, etc (Fletcher, 2016). Parfit's objective list includes "moral goodness", "rational activity", "development of abilities", "having children and being a good parent", "knowledge", and "the awareness of true beauty" (Parfit, 1984). Finnis developed a list that consists of "life", "knowledge", "play", "aesthetic experience", "sociability (friendship)", "practical reasonableness", and "religion" (Finnis, 2011). Murphy proposed a list with components of "life", "knowledge", "aesthetic experience", "excellence in play and work", "excellence in agency", "inner peace", "friendship and community", "religion", and "happiness" (Murphy, 2001). Finally, Fletcher's objective list includes the elements of "achievement", "friendship", "happiness", "pleasure", "self-respect", and "virtue" (Fletcher, 2013).

1.1.1.4. Desire fulfilment (or preference satisfaction) theory

The idea of this theory is that the well-being of an individual depends largely on the extent that his own desires are met (Brey, 2012; Hooker, 2015). In other words, the value of life improves when preferences are satisfied. The emergence of welfare economics was rooted in this theory in which a utility function is measured by the level of satisfaction.

1.1.1.5. Economic well-being

Traditional economic theory assesses the well-being of an individual based on the level of satisfaction (utility) obtained by the amount of goods and services consumed. The policy goal of an economy relies on the maximisation of social indicators such as gross domestic product (GDP) per capita, which is used to measure the average level of production of a country within a specified time period. However, the development of the paradox of happiness urged the use of happiness indices such as gross national happiness to measure economic achievement, due to increasing evidence regarding the lack of strong correlation between income or wealth and human satisfaction (Easterlin, 1974; Scitovsky).

1.1.1.6. Capability theory

This theory is based on a core concept that well-being is not purely determined by the amount of goods and services possessed by an individual. Specifically, whether one has the *ability* to execute the functions of those goods and services and other non-utility characteristics should be taken into consideration. This theory will be discussed comprehensively in section 2.2.3 of Chapter 2.

Besides the limitations of the widely used instruments to derive QALYs in terms of their restrictive focus on certain health-related dimensions without capturing other aspects of well-being, there is an increasing emphasis on mental well-being as it has been shown to be correlated with many aspects of morbidity, mortality and community outcomes (Barry *et al.*, 2009; Barry, 2009; Chida & Steptoe, 2008; Davidson, 2004; DiMatteo *et al.*, 2000; Friedli & Organization, 2009; Huppert & Baylis, 2004; Lyubomirsky *et al.*, 2005; Pressman & Cohen, 2005; Steptoe *et al.*, 2005). The promotion of mental well-being has therefore developed into a central goal for public health agencies tasked with delivering promotion activities and related interventions competing for limited publicly funded healthcare resources. Consequently, challenges to the use of QALYs in economic evaluation have been raised as they tend to underestimate the effect of interventions aimed at improving mental well-being. An alternative outcome measure called the “*Well-being Adjusted Life Year*” (I will use the term “*Mental Well-being Adjusted Life Year (MWALY)*” throughout this thesis instead due to its focus on mental aspects of well-being) has been formally proposed as an alternative to existing outcome measures with the view to capturing the benefits of interventions that focus on mental well-being (Johnson *et al.*, 2016). However, there is currently no preference-based tariff (valuation set) available for any pure mental well-being scale. Instead of focusing on certain “health” constructs, the adjustment component to life years within the MWALY aims to capture aspects of mental well-being.

Moreover, interest in the capability approach, originally proposed in the 20th century (Sen, 1993), has increased and generated additional evaluative tools for potential application within economic evaluation, revealing the shortcomings of more widely used valuation instruments to derive outcome measures. In the context of ongoing debate over the identification of appropriate outcome measures in economic evaluation, it is worthwhile exploring alternative evaluative measures that can inform publicly funded healthcare

based decision-making, as well as broader methodological debates within the health economics discipline.

1.2. Objective and research questions

The aim of this thesis is to develop a preliminary preference-based tariff for a mental well-being scale, in order to allow calculation of MWALYs. Specifically, the following research questions surrounding the background information and methodological considerations to achieve this aim will be addressed in this thesis:

- *Research question 1:* Do any existing preference-based measurement approaches and instruments value mental well-being?
- *Research question 2:* Are there any mental well-being measures that can be used to develop a preference-based tariff?
- *Research question 3:* What is the best choice of instrument for the elicitation of a preference-based tariff to allow the calculation of MWALY?
- *Research question 4:* What is the appropriate valuation protocol for the valuation of mental well-being state?

1.3. Structure of thesis

The following chapters were written to answer the above research questions.

Chapter 2 answers the research questions 1-3 by firstly reviewing the theoretical concepts of preference-based measurement approaches. This was followed by comparing existing generic preference-based instruments to reveal their constraints on valuing mental well-being. Next, existing non-preference-based mental well-being instruments were compared to justify the best choice of instrument for preference elicitation.

The research question 4 is answered in Chapters 3-6. Chapter 3 documents the methodology used in this thesis by reviewing a six-stage method in developing a preference-based instrument (Brazier *et al.*, 2012). A comparative analysis between existing valuation techniques was produced to propose a mental well-being valuation protocol. Afterwards, an overview of the two piloting phases (i.e. a qualitative phase and then a quantitative phase) for analysing the validity of the proposed valuation protocol through collecting primary data of valuation responses in the UK was provided. Chapter 4 documents the result of the qualitative phase, which was a think-aloud study to

investigate the participants' cognitive process of completing the valuation tasks. The results were used to improve the proposed valuation protocol. Chapters 5 and 6 document the result of the quantitative phase, which explored further the validity of the modified valuation protocol informed by the qualitative phase within a larger UK sample. Chapter 5 presents the feasibility, practicality and face validity of the participants' valuation responses. Chapter 6 applies econometric modelling techniques to derive preliminary versions of a preference-based valuation set.

Finally, Chapter 7 integrates the results from the qualitative and quantitative phases and discusses the implications of the main result. The chapter ends by discussing the contribution of this research, application of the valuation sets in economic evaluations and some future research directions.

Chapter 2: Review of measurement instruments

2.1. Introduction

In this chapter, the theoretical concepts underpinning alternative approaches to preference-based outcomes measurement that can guide healthcare decision-making, including the welfarist approach and variants of the extra-welfarist approach will be reviewed (section 2.2). After the construction of a solid theoretical foundation, the characteristics and differences of existing preference-based instruments within each theoretical approach that have been widely-used across the world to inform healthcare decision-making will be described (section 2.3). The comparative result of the identified instruments in terms of their coverage of constructs related to physical and mental health dimensions will be used to inform their abilities and roles in the valuation of mental well-being. Next, a comparative analysis of non-preference-based mental well-being instruments will be conducted (section 2.4). The results obtained from sections 2.3 and 2.4 will be synthesised in sections 2.5 and 2.6 to figure out the best choice of instrument for preference elicitation, so as to allow estimation of MWALYs. Section 2.7 concludes this chapter.

2.2. Theoretical concepts of welfarism and extra-welfarism

The theoretical concepts and main differences between the welfarist and extra-welfarist approaches in the context of healthcare decision making are reviewed in this section and are summarised in Table 1. It is noted that there are overlapping ideas between these approaches and their theoretical foundations are not completely separable from each other.

Table 1: Comparison of the concepts of welfarism and extra-welfarism

| | Welfarism | Extra-welfarism |
|---------------------------------------|--|---|
| Components of the objective function | Individual utilities, which are measured by the total amount of goods and services possessed. | Individual utilities, supplemented by a greater focus on other indicators of well-being including health, capabilities and functionings, freedom to choose, quality of relationship between individuals, etc. |
| Source of valuation of outcomes | The individuals affected by the interventions. | A representative sample of the general population is used to generate societal values. |
| Weighting of outcomes | Low concern for equity in which weighting and the distribution of welfare are unimportant. Weighting is sometimes still possible in a social welfare function. | High concern for equity and distribution of outcomes. Weighting is important such as equalising health and reducing health inequalities amongst the general public. |
| Interpersonal comparisons of outcomes | Impossible or meaningless, although it is sometimes possible within a social welfare function. | Meaningful and it is not restricted to the comparison of individual utility. Examples include the comparison of QALYs and ability to achieve something, etc. |

Sources: Brouwer *et al.* (2008), Coast *et al.* (2008b), Coast (2009), Coast *et al.* (2008c), Seixas (2017), Burchardt and Hick (2017), Birch and Donaldson (2003).

QALY indicates Quality-Adjusted Life Year.

2.2.1. Welfarism

The concept of welfarism has been incorporated into the economic evaluation of healthcare interventions in terms of informing approaches to resource allocation (Robin W. Boadway & Bruce, 1984). The fundamental principle of welfarist economics is that individuals are the best assessors of their own welfare and societal objectives should focus on the maximisation of individual utilities, in which individual utility is a function of an

overall sum of consumer goods and services (Seixas, 2017). The higher the amount of goods and services consumed and preferred by an individual, the higher would be the level of satisfaction or enjoyment gained by the individual himself or herself, leading to a higher utility outcome. Moreover, the judgment of the level of social welfare is based on the Pareto principle (Johansson, 1991). This states that an improvement in social welfare is achieved when an increase in utility for an individual does not cause a utility loss for another individual, *ceteris paribus*. Moreover, it is also beneficial and worthwhile to reach a particular state when the gain of individuals resulting from the movement to that state can compensate the individuals who suffer from this movement. In terms of weighting of utility (benefit), welfarism generally regards the distribution of welfare as irrelevant in decision-making as it has low concern for equity (Brazier *et al.*, 2017). However, weighting is still possible as some empirical attempts at the construction of a social welfare function have proposed assigning weights to the distribution of individual utilities and allowing interpersonal comparison of utility outcomes even though traditional welfarism treats the comparison of utility between individuals as meaningless and inapplicable (Brouwer *et al.*, 2008; Burk, 1938; Samuelson, 1947).

2.2.1.1. Application in economic evaluation

Economists conducting economic evaluations generally apply the welfarist approach in the form of cost-benefit analyses. Within this form of economic analysis, consequences or net benefits of an intervention are valued in monetary terms by consulting an individual's willingness to pay for the gain associated with that intervention. The human capital approach is sometimes also used to value benefits in which the value of an intervention or programme is calculated by the present value of incremental lifetime earnings associated with the gain from that intervention or programme as a mean of measuring productivity changes (Brent, 2014). In addition, cost-effectiveness analysis using QALYs as the measurement of health benefit is also performed by a few groups of welfarist proponents (Brazier *et al.*, 2017; Buchanan & Wordsworth, 2015). In this approach to economic evaluation, QALYs represent the utility level of an individual over health status. Particularly, in order to quantify preferences for individual patients in terms of years of life and health status, three properties of preferences are of paramount importance: utility independence, constant proportional trade-offs, and risk neutrality (Pliskin *et al.*, 1980).

2.2.2. Extra-welfarism

The concept of extra-welfarism (sometimes referred to as non-welfarism) in health economics was firstly promoted by Anthony John Culyer (Culyer, 1991; Culyer, 2012). Contrary to the theory of welfarism in which individual utility is a central source of “maximand” within the objective utility function, extra-welfarism expands the evaluative space from a restrictive composition of individual utility towards different types of non-utility or non-goods such as whether individuals have freedom to choose, capabilities to achieve something and their health status, etc. In other words, the maximisation of social welfare does not solely depend on the individual consumption of goods and services, but also on different indicators of well-being (Birch & Donaldson, 2003; Coast, 2009). It is worth noting that extra-welfarism does not completely exclude the importance of investigating the value of individual utility. Instead, it supplements the focus of valuing individual preferences by incorporating community preferences as a whole in the sense that individuals are not the only source of valuation of outcomes (Brouwer *et al.*, 2008). Additionally, extra-welfarism allows movement beyond the Pareto principle when determining the level of social welfare. Interpersonal comparisons of well-being in terms of capabilities and non-utility characteristics are permissible. There can be an improvement in social welfare by invoking different ethical criteria that may not necessarily be preference-based even if the beneficiaries do not compensate the sufferers. Weighting of relevant outcomes such as health production is crucial in the sense that extra-welfarism can incorporate concerns about equity. As extra-welfarism accommodates a wider evaluative space, weights are assigned under a broader perception of wealth, need and other ethical rules (Culyer, 1995; Culyer, 2007).

2.2.2.1. Application in economic evaluation

The concept of the extra-welfarist approach is mainly applied in cost-effectiveness analysis or cost-utility analysis. As proponents of extra-welfarism generally advocate the maximisation of health, the valuation of health states is not solely based on the affected individuals. A representative sample of citizens in society becomes the core assessors for the purposes of generating values for QALYs. In addition, non-welfarist cost-benefit analysis is also possible in economic evaluations. As opposed to the welfarist approach, non-welfarist monetary measurement and valuation of health benefits do not depend entirely on *individuals'* willingness to pay. Instead, the value of benefits can be evaluated

in terms of gains to the *citizen* after the allocation of public resources to alternative health interventions (Brazier *et al.*, 2017).

2.2.3. The capability approach

The theory of the capability approach was first introduced by the economist Amartya Sen (Sen, 1993), who has contributed to the investigation of human development theory in economic sciences. This approach provides a basis for the development of extra-welfarism in the sense that Sen challenged the inadequate focus of evaluating social welfare in terms of individual's preference-based utility only (Sen, 1979). In order to assess well-being or quality of life in broader terms, Sen explored the evaluative space through the inclusion of "functionings" and "capability" (Coast *et al.*, 2008c; Karimi *et al.*, 2016). Functionings are states and things someone has to be or to do, such as being happy, achieving self-respect, and being in good health. Capability represents a set of combinations of functionings that an individual can freely to choose from. The core value of this approach is that it is important to focus on whether a person is "able" to achieve something, instead of counting the amount of goods and services possessed by an individual. This point can be illustrated by the example of a private car. The possession of a private car is value-added to those living in a remote area where transportation is inconvenient or unavailable as it serves as an alternative way to increase mobility across that area. However, whether the functioning of being able to increase mobility can be achieved will depend on the personal conversion factors of an individual as these will affect the ability to convert the characteristics of the commodity into a functioning (Robeyns, 2005). An individual who has broken legs might not take this advantage as the person is not able to execute the driving action. Also, a healthcare specific example can be provided in this context. Malnutrition can be caused by fasting or a problem absorbing nutrients from food. The outcome of malnutrition is the same but the person who *chooses* to fast has a higher capability of changing the outcome whilst it is harder or impossible for the person to get rid of the nutrient absorption problem within a short period of time due to unfavourable underpinning health conditions. Considering this, a more accurate representation of true benefits can be achieved when individual ability is taken into account.

In addition, the capability approach puts a high concern on equity in which Sen discussed the concept of "capability equality" (Coast *et al.*, 2008b; Sen, 1980; Sen, 1982). Instead of aiming to maximise a certain objective function like the approaches of welfarism and

some extra-welfarist approaches, the capability approach is interested in investigating the gap between people with different levels of capability sets as it determines the extent of freedom to choose across combinations of functionings (Coast *et al.*, 2008b). It offers a new insight to the study of distributional or vertical inequality by analysing the distributions of functionings and capabilities, apart from the economic focus on income and wealth (Burchardt & Hick, 2017).

2.2.4. Summary

As a short summary, welfarism and extra-welfarism differ mainly in terms of the dimensions of the evaluative space that they cover. Welfarism advocates the improvement of social welfare through the maximisation of individual utility and it follows the Pareto and compensation principles in decision making. Weighting and interpersonal comparability of outcomes are unimportant or irrelevant. Under extra-welfarism, instead of focusing on the individual's preferences, broader measures of well-being and non-utility characteristics are taken into consideration such as health, capabilities and functionings. A representative sample of the public becomes the central source of valuation of outcomes. Weighting and interpersonal comparability of outcomes are allowed and there is a stronger advocacy for equity in the decision-making approaches. The capability approach falls under extra-welfarism in which it advocates the incorporation of individual abilities and functionings within the evaluative space and the restrictive focus on commodities and wealth to determine the well-being of individuals is claimed to be insufficient and inappropriate. Finally, it is worth noting that decision makers need to be cautious in their choice of type of economic evaluation approach as the adoption of different selected approaches can result in different policy decisions, due to the conflicting nature of the theoretical foundation behind different approaches (Buchanan & Wordsworth, 2015).

2.3. A comparative analysis of existing generic preference-based measurement instruments adopted across the world

After distinguishing the mainstream underpinning economic theories to healthcare decision-making, this section discusses the existing preference-based instruments used in different countries and jurisdictions for outcome measurement within the health economics literature. It will be followed by conducting a comparison regarding different

concepts and constructs covered within the questions of these instruments, so as to explore their differences and limitations in detail.

2.3.1. Description of preference-based instruments

Preference-based instruments can be classified into two categories based on the theoretical background behind their development: multi-attribute utility (MAU) measures and preference-based capability measures. All the measures in both categories fall under the extra-welfarist umbrella as they offer supplementary evaluative spaces that include health, different aspects of well-being and personal capabilities and functioning apart from individual consumption of goods and services within the objective function of decision makers.

2.3.1.1. MAU measures

The MAU instruments are generic measures generally used for deriving preference-based values for the health-related quality of life component of the QALY. They were developed under the framework of MAU theory in which at least first-order utility independence among the attributes should be met (Keeney & Raiffa, 1993; Michael F. Drummond *et al.*, 2015).

The variants of MAU instruments discussed in this report are listed below. Descriptions of these instruments are provided in Appendix 1.

- **Quality of Well-Being (QWB) scale**
 - *The Quality of Well-Being Scale Self-Administered (QWB-SA)*
- **The EuroQol five dimension measure (EQ-5D)**
 - *EQ-5D 3-level version (EQ-5D-3L)*
 - *EQ-5D 5-level version (EQ-5D-5L)*
- **Health Utilities Index (HUI)**
 - *HUI Mark 1 (HUI1)*
 - *HUI Mark 2 (HUI2)*
 - *HUI Mark 3 (HUI3)*
- **The Assessment of Quality of Life (AQoL) instrument**
 - *AQoL-4D*
 - *AQoL-6D*
 - *AQoL-7D*

- *AQoL-8D*
- **Short-Form Six-Dimension (SF-6D)**
- **The 15-dimensional (15D)**
- **Recovering Quality of Life - Utility Index (ReQoL-UI)**

2.3.1.2. Preference-based capability measures

The preference-based capability instruments discussed in this report are listed below.

Descriptions of these instruments are provided in Appendix 2.

- **Investigating Choice Experiments Capability Measure for Adults (ICECAP-A)**
- **Investigating Choice Experiments Capability Measure for Older people (ICECAP-O)**
- **The Adult Social Care Outcomes Toolkit (ASCOT)**

2.3.2. Concepts and constructs covered within the questions of the identified preference-based instruments

Table 2 provides a summary of the latest version of each of the instruments discussed in this chapter in terms of their preference-based nature, coverage of physical and mental health dimensions, and bases within major well-being theories.

Table 2: An overview of the concepts covered by the MAU and preference-based capability instruments

| | <u>QWB-SA</u> | <u>EQ-5D-5L</u> | <u>HUI3</u> | <u>AQoL-8D</u> | <u>SF-6D</u> | <u>15D</u> | <u>ReQoL-UI</u> | <u>ICECAP-A</u> | <u>ICECAP-O</u> | <u>ASCOT</u> |
|--|---------------|-----------------|--------------|----------------|--------------|---------------|-----------------|-----------------|-----------------|--------------|
| Number of dimensions/items/domains/attributes | 4 dimensions | 5 dimensions | 8 attributes | 8 dimensions | 6 dimensions | 15 attributes | 10/20 items | 5 attributes | 5 attributes | 8 domains |
| Health dimensions: | | | | | | | | | | |
| <i>Physical</i> | | | | | | | | | | |
| Senses | | | | | | | | | | |
| Vision | ✓ | | ✓ | | | ✓ | | | | |
| Hearing | ✓ | | ✓ | ✓ | | ✓ | | | | |
| Speech | ✓ | | ✓ | ✓ | | ✓ | | | | |
| Breathing | ✓ | | | | | ✓ | | | | |
| Eating | ✓ | | | | | ✓ | | | | |
| Elimination | ✓ | | | | | ✓ | | | | |
| Mobility/Activities | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| Self-care | ✓ | ✓ | | ✓ | | | ✓ | | | |
| Dexterity | ✓ | | ✓ | | | | | | | |
| Energy | ✓ | | | ✓ | ✓ | ✓ | | | | |
| Pain | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| <i>Mental</i> | | | | | | | | | | |
| Positive | | | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ |
| Negative | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| Hedonic view of well-being | | | | | | | | | | |

| | | | | | | | | | | |
|--------------------------------------|---|---|---|---|---|---|---|---|---|---|
| Positive affect | | | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ |
| Negative affect | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | |
| Eudaimonic view of well-being | ✓ | | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| Capabilities focus | | | | | | | | ✓ | ✓ | ✓ |

AQoL-8D indicates Assessment of Quality of Life - 8D; ASCOT, Adult Social Care Outcomes Toolkit; 15D, 15-dimensional instrument; EQ-5D-5L, EuroQol five dimension 5-level version; HUI3, Health Utilities Index Mark 3; ICECAP-A, Investigating Choice Experiments Capability Measure for Adults; ICECAP-O, Investigating Choice Experiments Capability Measure for Older people; QWB-SA, Quality of Well-Being Scale Self-Administered; ReQoL, Recovering Quality of Life; SF-6D, Short-Form Six-Dimension.

2.3.2.1. Comments on the coverage of physical health dimensions

At first glance at this table, it is obvious that *only* MAU instruments cover dimensions of physical health. One of the common characteristics for all MAU instruments is that questions related to the physical dimension of “Mobility/Activities” and “Pain” are included. HUI3 has a unique coverage of “dexterity”, which is about the ability of the use of hands or fingers with or without the help from another person or special tools. The QWB-SA provides the most comprehensive physical health coverage, which is expected due to its lengthy nature with the largest number of items regarding physical health. Moreover, the 15D provides the second most comprehensive coverage of physical dimensions. The attributes “breathing” and “eating” are independently classified as extra dimensions, relative to the other MAU instruments except the QWB-SA. Also, the 15D is the only instrument which has a specific attribute of “sexual activity” assessing whether a person’s state of health has any effects on their sexual activity. ReQoL only has one physical health item related to pain, mobility, self-care and feeling unwell.

2.3.2.2. Comments on the coverage of mental health dimensions

Regarding the dimensions of mental health, the questions in all of the MAU and preference-based capability instruments respectively cover different constructs of mental health. However, the extent of coverage varies a lot between these instruments. Generally speaking, the design of the MAU questions places less emphasis on assessing the mental health of the respondents, relative to the preference-based capability measures. Among all the MAU instruments, the EQ-5D-5L and the HUI3 have the *least* number of attributes addressing mental health and broader well-being concepts. The ReQoL has the highest proportion of mental health attributes included within the survey, due to its focus on assessing the recovery outcome of mental health service users. Whilst most of the MAU instruments contain only a tiny proportion of mental health items, published evidence related to constructs of mental health will also be highlighted in section 2.3.2.2.1 to ease comparison and supplement the analysis of the MAU instruments in the context of mental health dimensions.

2.3.2.2.1. *MAU instruments - Constructs and their applications or coverage of mental health dimensions*

2.3.2.2.1.1. QWB-SA

Firstly, for the QWB-SA, only 14 mental symptoms or behaviours are included in the descriptive system, occupying the lowest proportion compared to the coverage of all other physical health symptoms. The description of mental health is negatively worded, as indicated by the adjectives “upset”, “anxiety” and “lonely”, etc. There is only one statement mentioning control over events in one’s life, revealing an extremely low coverage of eudemonic well-being.

In terms of applications in the mental health context, the QWB-SA has been mainly used to assess mental illness. For example, Pyne *et al.* (2003) recruited patients with unipolar or bipolar depression to investigate the cross-sectional and longitudinal relationships between the QWB-SA and Hamilton Rating Scale for Depression (HRSD-17) and Beck Depression Inventory (BDI). After following the patients with the treatment of antidepressant and mood stabilising medication, the rating score of the instruments revealed a statistically significant and negatively correlated relationship between QWB-SA and depression severity. In addition, another study analysed the minimal clinically important difference for the EQ-5D and the QWB-SA in post-traumatic stress disorder within the context of cognitive behavioural therapy and pharmacotherapy (Le *et al.*, 2013). Under the comparison of before and after the treatment in a preference trial, the ranges of utility scores for both instruments have been used to inform mental health treatment interventions.

2.3.2.2.1.2. EQ-5D-5L

For the EQ-5D-5L, only one out of five dimensions (anxiety/depression) relates to (negative) mental health of the respondents. Because of this, it is not surprising to discover that there is plenty of evidence suggesting its insensitivity in capturing the effects of interventions that improve mental well-being under the widespread use of the EQ-5D in health technology assessment (Brazier, 2010; Saarni *et al.*, 2010; Shah *et al.*, 2017a).

Saarni *et al.* (2010) investigated the decrement of subjective quality of life (QoL) and preference-based HRQoL associated with different kinds of psychotic disorders. Data were obtained from the Health 2000 survey, which incorporated a sample of 8028 members of the Finnish population following screening for psychotic disorders. Subjective QoL was assessed using a VAS within the range of 0-10, whereas the EQ-5D was used as one of the preference-based HRQoL measures. Multiple regression models

were used to estimate the relationship between psychotic disorders and the reduction of QoL or HRQoL. The main results showed that the EQ-5D failed to detect statistically significant decrements in HRQoL associated with delusional or bipolar I disorder, after controlling for age, gender and other socio-demographic factors. In addition, Shah *et al.* (2017a) investigated the views of the UK general public on important aspects of health that were absent from the EQ-5D instrument. Face-to-face interviews among a sample of the UK population were conducted to gather views from respondents regarding the dimensions of health that are not captured by the EQ-5D. Among the 436 included respondents, around 40% of them agreed that there were some missing themes within the EQ-5D. General mental health and some specific conditions and disorders that affect cognitive functioning such as stress and dementia, were regarded as important missing aspects identified by the respondents. In the context of this evidence, although the EQ-5D descriptive system includes a dimension that assesses the level of anxiety or depression, it may merely reflect mild to moderate mental health conditions and is considered by some to be inadequate for use in psychotic disorders (Brazier, 2010).

Lastly, it is argued that the long-term reliability of the results obtained from the EQ-5D is questionable as the questionnaire only focuses on the description of health “today”. The mental health condition of an individual can fluctuate from time to time and may not be fully captured in the EQ-5D results. In this sense, the value of effective mental health interventions may be underestimated.

2.3.2.2.1.3. HUI3

For the HUI3, one out of eight attributes named “emotion” addresses concepts other than physical health, and this attribute attempts to capture more constructs of mental health and well-being theories, relative to the EQ-5D-5L. The description of this attribute is ranged from level 1 (Happy and interested in life) to level 5 (So unhappy that life is not worthwhile). The use of the adjectives “happy” and “unhappy” reflect the conceptual coverage of positive and negative affects, so as to assess the mental health condition of the respondents. The concept of eudemonic well-being is partially covered by this attribute in the sense that it also assesses whether respondents can realise the meaning and value of life through their attitudes towards life.

Due to the restrictive mental health focus of the HUI3, researchers mostly investigated this instrument in terms of emotion-based mental health construct. For example, Luo *et al.* (2006) investigated the construct validity of the HUI3 in patients with schizophrenia, which is a mental disorder. Patients were recruited from a tertiary mental hospital in Singapore to complete the HUI3, the Short-Form 36 Health Survey (SF-36) and the Schizophrenia Quality of Life Scale (SQLS). Through a statistical estimation of the correlation between HUI3 and other instruments, the empirical results showed that there were statistically moderate correlations between the “emotion” attribute in the HUI3 and the “mental health” attribute in the SF-36 (Spearman’s rank correlation = 0.45) and the SQLS psychosocial scales (Spearman’s rank correlation = -0.43). Also, the mean score for the “emotion” attribute in the HUI3 was 0.09 lower than the population norms adjusted for sex, age and ethnicity. This suggested a discriminatory power and an acceptable construct validity for the HUI3 in the context of schizophrenia. In addition, Chu *et al.* (2017) discovered the relationship between mental health and the serum 25-Hydroxyvitamin D concentrations based on the data obtained by the Canadian Health Measures Survey. The “emotion” attribute of the HUI3 was used as one of the proxy measures for assessing three indicators of mental health, which were depression, anxiety and stress. The result from the ordered logistic regression models robustly showed that there was a positive association between serum 25-Hydroxyvitamin D concentrations and being in the best emotional health category of the HUI3. Over and above that, there is a study aiming to empirically analyse the construct validity of the HUI emotion scores and other mental health measures, including the non-specific psychological distress (K6) and the Composite International Diagnostic Interview (CIDI) (Feeny *et al.*, 2009). Data were obtained from the Statistics Canada National Population Health Survey Cycle 2, which was targeted at persons aged 12 or above. The correlation coefficients of -0.46 and -0.31 were observed between HUI emotion scores and the K6, and between HUI emotion scores and the CIDI respectively. These medium levels of correlation provided evidence of the construct validity of the HUI3 emotion category in the assessment of population mental health.

2.3.2.2.1.4. SF-6D

The SF-6D contains an attribute assessing respondents’ level of negative mental health. This attribute is negatively worded and is ranged from level 1 to level 5. Moreover, there

is coverage of eudemonic well-being in this instrument as it also uses the frequency of social activities as an indication of social functioning.

There are several articles comparing the SF-6D with other measures in the context of mental health. For example, Lamers *et al.* (2006) used the data from a multicentre randomised trial in Mental Health Care Centres in the Netherlands to compare EQ-5D and SF-6D utilities in patients with mood and/or anxiety disorders. The discriminatory ability of the EQ-5D and SF-6D between subgroups of patients with different levels of disorder severity was also examined. Based on the result of the Spearman's rank correlation between EQ-5D and SF-6D dimensions, a positive coefficient of 0.415 was found between the "mental health" dimension of the SF-6D and the "Anxiety/depression" dimension of the EQ-5D, which was the second strongest relationship among all the dimension correlations. This indicates a moderate strength of correlation between the coverage of mental health constructs between these two instruments. Relative to the SF-6D, although a higher proportion of respondents reported the EQ-5D scores at the upper end of the scale (no problems) and a lower proportion of them reported at the lowest end of the scale (extreme problems) in general, there was an exceptionally different finding for the mental functioning dimension in both instruments. For the dimension of "anxiety /depression" within the EQ-5D instrument, more than 33% of the patients reported extreme problems, a much higher proportion when compared to the same level in the other four dimensions with reported percentages from around 1% to 10%. For the "mental health" dimension of the SF-6D instrument, around 65% of patients reported the lower end of the scale (levels 5 and 6 representing the feeling of depression and nervousness most and all of the time respectively), which was the highest proportion of lower level responses of all dimensions. This is intuitive as the patients had mental disorders. However, the interesting finding here is that there is a great variation (65% versus 33%) between the instruments in terms of the sensitivity in detecting the level of mental functioning. Moreover, the statistically significant difference in mean utilities among all four adjacent subgroups categorised by the quartiles of the Symptom Checklist (SCL-90) for both EQ-5D and SF-6D instruments revealed the discriminatory ability of severity in mental problems. Lastly, those subgroup with the most severe mental health problems demonstrated a larger improvement in mean utilities for the EQ-5D from baseline to 1 and 1.5 years follow-up, relative to the mean improvement for the SF-6D. Decision makers should be cautious in the choice of instrument due to their differences in detecting patients with severe mental health

problems, resulting in different cost-utility ratios when conducting economic evaluation. In addition, Bharmal and Thomas (2006) assessed the ceiling effects of the EQ-5D and SF-6D in the US population based on the data obtained from the 2000 Medical Expenditure Panel Survey. Interestingly, the results showed that 47% and 5.8% of respondents reported full health across all of the EQ-5D dimensions and SF-6D dimensions, respectively. Also, around 49% of the respondents who reported no limitations (full health) in the EQ-5D were actually classified in level 2 or above in the SF-6D mental health dimension, demonstrating some problems regarding the frequency of suffering negative mental health. Indisputably, the EQ-5D was limited by its substantial ceiling effect. Even though a 5-level version of the EQ-5D has been developed to overcome the problem of its ceiling effect, future research will be required to investigate its performance in capturing changes in mental health conditions when compared with the SF-6D and other instruments.

2.3.2.2.1.5. 15D

Although more than half of the attributes of 15D are related to physical health, it has several attributes covering the mental health of respondents. For example, the attributes “depression” and “distress” are described by negatively worded adjectives such as “sad”, “melancholic” and “anxious” to investigate negative aspects of mental health. Also, there is a specific attribute of “mental function” to assess the memory and thinking of the respondents. These mental health dimensions are used to increase the discriminatory power and responsiveness to change in health status of individuals (Sintonen, 2001).

There are several articles discussing the 15D in the context of mental health. For example, Anagnostopoulos *et al.* (2013) empirically examined psychometric and factor analytic properties of the 15D within the Greek general population. Two samples were extracted in which one was for exploring the distributional properties and factor structure of the 15D and another one was for its use in confirmatory factor analysis. Among all the statistical results, one of the crucial findings was that among the three factors of the 15D items in the Promax rotated factor matrix (Factor I: functional ability; Factor II: physiological needs satisfaction; Factor III: emotional well-being), the factor of emotional well-being was the only one indicating discriminant validity, even though convergent validity was supported for all three factors. However, in addition to that, the attribute of “mental function” demonstrated a substantial ceiling effect. Specifically, 96% of the respondents

selected the best response options (level 1) in this attribute, which was the highest among all other 15D attributes. The best response option of the “depression” attribute was also reported by around 74% of respondents. These suggested the limitations of these attributes in reflecting the mental health condition of the general population sample effectively. Additionally, Leppanen *et al.* (2016) analysed the most applicable model for patients with severe borderline personality disorder (BPD). The 15D was used as an instrument to assess the HRQoL of the BPD patients and the 15D scores were compared with age-matched general Finnish population scores. The result showed that there were statistically significant lower mean values for the 15D dimensions related to “mental function”, “depression” and “distress”, relative to the general population. After one year of trial implemented through the randomisation of the patients into two groups named the Community Treatment by Experts (CTBE) and the Treatment as Usual (TAU), the CTBE group revealed a substantial improvement of the quality of life and reduction of BPD symptoms, relative to those of the TAU patients. These results provided important implications for the choice of treatment model.

2.3.2.2.1.6. AQoL-8D

There are 35 questions in total within the AQoL-8D instrument with comprehensive coverage of mental health and well-being dimensions. In terms of positive affect, it has a number of questions that explicitly ask the frequency or extent of feeling happy, pleasure and enthusiasm, etc. Questions regarding the assessment of negative affect such as the frequency of feeling angry, feeling depressed and despair are also included. Furthermore, the AQoL-8D has a higher proportion of questions addressing eudaimonic well-being, compared to the other MAUs. For instance, questions regarding the level of confidence possessed by the respondents and the frequency of feeling worthless allows the interpretation of self-acceptance and human value. Questions regarding the frequency of feeling socially isolated and the satisfaction of relationships with family and friends offer clues to assess the relationships of respondents with others. It is noteworthy that the AQoL-8D attempts to strike balance between the coverage of physical health, mental health and broader aspects of well-being within the questionnaire.

The most crucial work discussing the coverage of the mental health dimensions of the AQoL is that which aimed to conduct a psychometric analysis for exploring the sensitivity of the AQoL-8D for the economic evaluation of interventions affecting mental health

(Richardson *et al.*, 2011a). By extending the inventory dimensions of the AQoL-6D, the AQoL-8D was shown to be more sensitive to the construct of mental health. By examining the statistical differences between the mean mental health patient and public values for all AQoL-8D dimensions, significant results were identified for every dimension. Specifically, the difference was largest for the AQoL-8D mental health dimension, reflecting its ability to detect mental health problems. Across the general population and patient groups with AQoL-8D scores less than AQoL-6D, a large magnitude of smaller mean values for the mental health dimensions and larger mean values for the standardised physical dimension scores were found, supporting a stronger sensitivity of the AQoL-8D in the dimensions of mental health. The validity of these results was further supported by the lack of ceiling effect of the AQoL-8D overall score (public: 0%; patient: 0.4%) and its corresponding mental health dimension score (public: 0%; patient: 1.2%), as indicated by the percentage of respondents reporting the maximum score in the questionnaire.

2.3.2.2.1.7. ReQoL-UI

Ten out of 11 ReQoL-10 items are related to mental health. Among these items, the hedonic well-being is assessed by the extent of individual happiness and loneliness, etc. The eudemonic well-being is assessed by, for instance, how does the individual think about the worthiness of life and the level of self-confidence. These are elements of reflecting the true value and meaning of life. The item of feeling able to trust others assesses the ability to engage in meaningful social contact with the others. For the ReQoL-20, twenty out of 21 items are related to mental health. In addition to the 10 mental health items included in the ReQoL-10, ReQoL-20 expands the coverage of hedonic well-being by assessing the irritated, terrified, anxious, and calm feelings. Also, eudemonic-related items such as doing rewarding things, feeling control of life, and the feeling of being a failure are additionally included to assess the individual's position or management of life.

The mental health application of these two versions of ReQoL is targeting individuals with mental health difficulties. For example, the testing of item development process and the validation of psychometric properties among these two surveys were built on data from service users, clinicians for mental health service providers, and patients in the UK (Keetharuth *et al.*, 2018a). The data from the general population was included as a comparator to investigate the sensitivity of ReQoL to distinguish between general population and service users. The result supported the ability of ReQoL to detect

differences between the general UK population and individuals with schizophrenia, bipolar, personality, psychotic and other common mental health disorders. Another study analysed the psychometric validation of the ReQoL-10 within a sample of first-episode psychosis intervention program in Singapore (Chua *et al.*, 2020). The result supported the internal consistency of ReQoL-10 (Cronbach's alpha = 0.89). It demonstrated a mixed performance on construct validity, when investigating the expected correlation relationship between the scores of ReQoL-10 and other rating scales such as PHQ-9, EQ-5D-3L, and the Positive and Negative Syndrome Scale at different time points of intervention. Also, it was interesting to realise that demographic and clinical characteristics regarding age, marital status, education level, duration of untreated psychosis and diagnosis of affective psychosis were associated with different ReQoL-10 scores.

2.3.2.2.2. *Preference-based capability instruments - Constructs of mental health dimensions*

It is noted that the capability measures, though preference-based, don't generate QALYs as they are not anchored at 0 (dead) to 1 (full health) scale. As mentioned before, none of the preference-based capability measures covers physical health and comparison between them will be based on their coverage of mental health constructs only.

2.3.2.2.2.1. ICECAP-A and ICECAP-O

For the ICECAP-A, there are attributes covering the positive affect of respondents, including the level of possession of enjoyment and pleasure, the feeling of love, etc. There is no negatively worded attribute for the assessment of negative affect. There is also coverage of eudemonic well-being as it has an attribute assessing whether respondents can achieve and progress any aspects of life. This can explore the capability of the respondents to identify their real potential and true self through their performance at different stages of life experience. Besides, the ICECAP-O also contains attributes assessing the enjoyment, pleasure and love that the respondents want. Coverage of eudemonic well-being can be shown by the attribute regarding whether respondents are able to do things that make themselves feel valued. All the response items within both the ICECAP-A and the ICECAP-O include capability wordings either ranged from "I am able to" to "I am unable to" or from "I can" to "I cannot". Both questionnaires contain a specific attribute

assessing the ability of the respondents to be independent, which is a unique feature relative to all other instruments discussed in this chapter.

2.3.2.2.2.2. ASCOT

The ASCOT offers minor coverage of positive mental health as there is a dimension assessing the respondents' level of feeling safe and whether this level fulfils their safety wants. There are also several domains covering eudaimonic well-being. For example, there is a domain assessing the level of social contact with the others and the domain level of "Dignity" investigates the extent to which the way respondents are helped and treated affects their thinking and feeling about themselves, revealing the level of self-acceptance and self-reflection. Different from the other capability measures, the ASCOT aims to investigate whether their wants and needs are fulfilled through the assessment of different aspects of social care. In this context, the description of domain levels is mostly related to whether certain social characteristics are satisfied such as the ability to get adequate or timely food and drink, the ability to enjoy their occupation tasks, and whether the desire of a clean and comfortable accommodation is met.

2.3.2.3. Overview of instruments with regard to measuring mental well-being

In considering the value of available preference-based instruments in the valuation of mental well-being, I have considered the following criteria:

- Coverage of the mental versus the physical components of health
- Coverage of well-being versus disease aspects of the mental components
- Aims of the developers for constructing the measure - to aid decision making in sick or well populations
- Validation in sick or well populations
- The theoretical underpinning of the scales

In terms of coverage of mental health constructs, most of the MAU instruments contain fewer attributes or dimensions related to mental health, relative to the physical health attributes and dimensions. QWA-SA, EQ-5D-5L, SF-6D and 15D only assess negative aspects of mental health, without evaluating mental well-being. ReQoL is the MAU instrument with the least number of physical health item. It aims to assess the mental health performance of mental health service users. HUI3 and AQoL-8D assess both positive and negative mental health. However, the HUI3 places much stronger emphasis

on physical health as seven out of eight attributes are assessing the physical health. Although the design of AQoL-8D aims to increase the coverage of mental health dimensions, the negatively worded questions occupy a significant proportion of the questionnaire. One should be cautious to regard it as an effective instrument in valuing positive mental health until there is published evidence discussing its ability to capture mental well-being benefits, when compared with instruments with a pure mental well-being focus.

For the preference-based capability measures, it is encouraging that they all cover positive aspects of mental health, but not negative aspects. However, I would argue that the role of all these measures is not tailored to measure mental well-being as they were developed based on capability theories. Regarding the nature of questions, the word phrasings used within capability measures are uniquely designed to reflect the individual's capabilities and functionings. Because of this, phrases such as "I am able to be", "I am unable to be", "I can" and "I cannot" frequently appear within capability measures.

Overall, although ReQoL focuses on the assessment of mental health, it is not a pure mental well-being instrument. It is not restricted to measuring positive spectrum of mental health, as it contains both positively worded and negatively worded mental health items and one physical health item. Also, it is not a generic measure for the completion of general population, but only generic in terms of measuring recovery-focused outcomes for service users with generic range of mental health problems or diagnosis. AQoL-8D and preference-based capability instruments can be regarded as offering potential candidates in the role of measuring mental well-being. However, they might map into mental well-being only partially and not be comprehensive at capturing the benefits of mental well-being interventions.

With a view to precisely estimate the MWALYs for the economic evaluation of interventions targeting at members of the general population, it is important to identify an instrument and its corresponding utility set with a stronger and comprehensive focus on generic mental well-being. Besides the aforementioned preference-based instruments, the next section will explore and compare existing non-preference-based mental well-being instruments, to identify the best candidate for the measurement of mental well-being.

2.4. A comparative analysis of non-preference-based mental well-being measures

This section reviews and compares the availability of mental well-being instruments, to inform the choices of instrument in accurately measuring mental well-being.

As there are numerous mental well-being measures available, inclusion criteria for measures to be reviewed in this section were developed. These were informed by a published systematic review of instruments measuring mental well-being (Rose *et al.*, 2017). The selected instruments in this report are based on the following inclusion criteria:

- ✓ Focus on mental health (i.e. more than 50% of the items are related to mental health dimension)
- ✓ Validated on general population samples for the measurement of mental well-being
- ✓ Positively worded for all items
- ✓ Contain at least one item that assesses feeling (hedonic) and one item that assesses functioning (eudemonic)
- ✓ Not only applied in childhood
- ✓ Widely used in the UK, which is the target country in this project

The mental well-being instruments discussed in this report are listed below. Descriptions of these instruments are provided in Appendix 3.

- **Warwick-Edinburgh Mental Wellbeing Scales (WEMWBS) and Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS)**
- **The World Health Organisation- Five Well-Being Index (WHO-5)**
- **Mental Health Continuum-Short Form (MHC-SF)**
- **Social and emotional health survey (SEHS)**

As the mental well-being instruments contain only positively worded items and do not cover any physical health dimensions, the comparison is concentrated on the assessment of mental well-being dimensions, addressing the limitations of most of the MAU and preference-based capability instruments in reflecting the benefits of mental well-being interventions in economic evaluation.

2.4.1. WEMWBS/SWEMWBS

WEMWBS and SWEMWBS both focus on the coverage of mental well-being and they differ in the number of items within the questionnaire. The frequency of the feeling of

positive mental health is addressed with adjectives such as “optimistic” and “relaxed”, etc. Besides the coverage of positive affect, items are designed to cover eudemonic well-being. For example, there are items mentioning the feeling of usefulness and the close feeling with other people. These are related to the realisation of true self and the relationship with others. Due to the nature of the assessment of positive mental health, WEMWBS and SWEMWBS are widely used in the evaluation of health promotion programmes or interventions. These will be summarised in section 2.6 to support the promotion of WEMWBS or SWEMWBS as a suitable basis for the development of an alternative preference-based outcome measure.

2.4.2. WHO-5

The WHO-5 is used to measure the positive aspect of mental health with one item specifically capturing the eudemonic well-being. Adjectives such as “cheerful”, “relaxed”, “active” are used within the items as a description of mental well-being. The item “My daily life has been filled with things that interest me” helps explore human potential by looking at whether respondents can frequently find meaning and motivation in life. Although both WHO-5 and WEMWBS/SWEMWBS aim to assess positive mental health, WEMWBS/SWEMWBS have a broader coverage of eudemonic well-being, as indicated by a higher number of items related to psychological functioning, social relationship and self-actualization.

2.4.3. MHC-SF

The MHC-SF is also used to measure the level of mental well-being of an individual. The positively worded adjectives such as feeling “happy” and “confident” describe the positive affect of hedonic well-being. Also, eudemonic well-being plays an important role in the classification system in the sense that items such as “your life has a sense of direction or meaning to it”, “you had warm and trusting relationships with others” and “you had something important to contribute to society” aim to assess individuals’ recognition of their positive relationships with the others, true values and meaningful of life. The main difference between MHC-SF and WEMWBS/SWEMWBS or WHO-5 lies in their coverage of hedonic and eudemonic items. Specifically, around 60%-80% and 80%-99% of the items are related to functioning for the MHC-SF and SWEMWBS respectively (Rose *et al.*, 2017). The items within the WHO-5 are predominantly (80%-99%) related to feeling. WEMWBS has a balanced coverage of both feeling and functioning.

2.4.4. SEHS

The SEHS aims to measure the mental well-being of youths studying between Grade 7 and Grade 12. There are items assessing hedonic well-being, such as “There is a feeling of togetherness in my family” and “Since yesterday how much have you felt grateful”. Different from the MHC-SF, WEMWBS and WHO-5 but similar to SWEMWBS, the items in SEHS are predominantly functioning, contributing 80% - 99% of the total items (Rose *et al.*, 2017). Examples of eudemonic items include “There is a purpose to my life”, “I understand why I do what I do” and “I can deal with being told no”, etc. These reflect the realisation of individual potential and the understanding of self-values. The main difference between SEHS and SWEMWBS is based on aspects of well-being coverage. Instead of mainly assessing an individual’s own mental well-being, the SEHS contains specific items evaluating the individual’s belief in others, such as “I have a friend my age who really cares about me” and “my family members really help and support one another”. Also, some of the items in SEHS are school-based in the sense that these are phrased as “At my school, there is a teacher”, with a view to assess individual’s belief in school teachers. Moreover, there is no fixed recall period for all items in SEHS. Mental well-being for specific items are assessed with different recall periods including “yesterday” and “right now”.

2.4.5. Summary

Although WEMWBS/SWEMWBS, WHO-5, MHC-SF, SEHS, ICECAP-A, ICECAP-O and ASCOT cover aspects of positive mental health, the corresponding proportion of coverage and the core focus of the question design are not identical. WEMWBS/SWEMWBS, WHO-5, MHC-SF and SEHS aim at measuring mental well-being, so all of the item descriptions are positively worded to reflect the assessment of positive mental states. However, as ICECAP-A, ICECAP-O and ASCOT aim to assess capabilities and functionings, there are relatively far fewer dimensions covering the concept of mental well-being. Undoubtedly, WEMWBS/SWEMWBS, WHO-5, MHC-SF and SEHS could be more sensitive at detecting the benefits of mental health promotion interventions.

Also, in terms of evaluating mental well-being or mental health promotion interventions, mental well-being instruments could be better than AQL-8D as the role of AQL-8D is not specifically tailored to measure pure mental well-being benefits due to its

incorporation of both physical and mental health questions. ReQoL is not designed for the completion by general population, even though a wide range of mental health items is included. As the goal of this project is to develop a preference-based tariff for an instrument used to specifically assess mental well-being among members of the general population, the contents of items within the WEMWBS/SWEMWBS, WHO-5, MHC-SF and SEHS seem more relevant in this sense.

2.5. Justification for the need to estimate the MWALY through the development of a preference-based tariff for a mental well-being instrument

The aforementioned sections provide an overview of all the common measurement instruments and a comparative analysis regarding their coverage of different constructs of physical and mental health has been performed. In this section, with the aid of the findings from the instrument comparison, the reasons for the development of an alternative outcome measure to QALY will be justified before moving on to discuss the best candidate for preference elicitation.

First and foremost, although QALY is recommended by many decision-making bodies such as the National Institute for Health and Care Excellence (NICE) as an outcome measure in economic evaluation (Whitehead & Ali, 2010), the widely used MAU instruments for the derivation of the HRQoL component of the QALY are subject to critiques and limitations due to their restrictive focus on certain “health” dimensions when measuring and valuing the benefits of an intervention. Specifically, as mentioned in the previous section, a large proportion of dimensions for most of the MAU instruments are related to physical health. Without a sufficient incorporation of dimensions related to mental health or broader aspects of well-being, there would be a substantial underestimation of the benefits related to well-being interventions when the QALY is adopted as a measurement tool. As there are many aspects of life affecting the living standard of an individual, people value things more than the absence of physical illness, revealing the limitations of QALY in the reflection of societal values. This notwithstanding, it is notable that AQoL-8D is one of the recent MAU instruments attempting to improve the degree of insensitivity to the construct of mental health by modifying the coverage of mental health dimensions in the questionnaire. However, the coverage emphasis of mental health and well-being components might not be as comprehensive as the other well-being or capability-based instruments, and there is still a

lack of evidence surrounding the comparison of the performance between AQoL-8D and other mental well-being or preference-based capability measures in the context of mental health and well-being interventions. ReQoL could be sensitive in capturing the change in outcome for individuals experiencing different forms of mental health problems, but its role is not orientated to target members of the general population without mental health issues.

In addition, there has been an increasing emphasis on the allocation of publicly funded healthcare resources to mental health promotion programmes as mental health status is shown to be crucially related to many aspects of disease and disability, and also broader aspects of social outcomes including educational attainment, labour productivity and earnings, etc (Barry *et al.*, 2009; Barry, 2009; Chida & Steptoe, 2008; Davidson, 2004; DiMatteo *et al.*, 2000; Friedli & Organization, 2009; Huppert & Baylis, 2004; Lyubomirsky *et al.*, 2005; Pressman & Cohen, 2005; Steptoe *et al.*, 2005). The promotion of mental well-being has therefore become an important healthcare goal, emphasising the need to develop an alternative outcome measure that accurately reflects the benefits of interventions related to mental well-being.

Lastly, given the importance of mental well-being nowadays, there is still currently no preference-based tariff for mental well-being measures. Although the capability instruments (ICECAP-A, ICECAP-O and ASCOT) discussed in this chapter are preference-based, they aim to assess aspects of capabilities and functionings. In light of this, there is no preference-based tariff for an instrument with a pure focus on mental well-being.

Based on the arguments presented above regarding the research gap in the construction of a preference-based tariff from a suitable well-being instrument, an alternative outcome measure is necessary to evaluate the benefits of an intervention beyond restrictive “health” dimensions. Considering this, the Mental Well-being Adjusted Life Year (MWALY) is formally proposed as an alternative outcome measure to the QALY to avoid the underestimation of significant mental well-being benefits of publicly funded healthcare interventions.

2.6. Discussion of the best choice of instrument for preference elicitation

When it comes to the elicitation of a preference-based tariff for a mental well-being instrument to allow the calculation of the MWALY, the WEMWBS/SWEMWBS could be the recommended choice of instrument given that it has received wide recognition across different sectors (Keyes, 2013).

Johnson *et al.* (2016) raised the need for the development of the MWALY, in order to overcome the imperfections of the QALY in reflecting the impacts of interventions that improve mental well-being. The possibility of mapping WEMWBS onto the EQ-5D-3L was analysed based on the data from the Coventry Household Survey, in which 7469 residents of Coventry aged 16 years or older were surveyed between 2011 and 2013. The results indicated no ceiling effect for the WEMWBS, and the WEMWBS scores of the participants in the survey were widely distributed across the range from 14 to 70. However, more than 70% of the participants who scored the maximum of 1.0 on the EQ-5D-3L index had a mean score of around 53.9 on the WEMWBS. Obviously, WEMWBS and EQ-5D-3L are not measuring identical aspects of health or well-being and the scoring of the EQ-5D-3L index is strongly limited by its ceiling effect. With the limited mapping relationship between WEMWBS and the EQ-5D-3L, a new outcome measure is required to reflect the benefits of interventions with a mental well-being focus.

Overall, service users of mental health care place a high rating on the WEMWBS in terms of its coverage of positive mental health. For example, Crawford *et al.* (2011) aimed to investigate the acceptability and relevance of widely used outcome measures in the view of secondary care mental health service users with experience of mood and/or psychosis disorders. Based on the contents of the measures, they were asked to judge the quality of the measures by performing an initial rating on an 11-point Likert scale. They were asked to perform a final rating after commenting and discussing the outcome measures and when feedback on the response results of the initial score were released. Measures of general mental health including the WEMWBS, the Clinical Outcomes in Routine Evaluation - Outcome Measure (CORE-OM) and the General Health Questionnaire-12 (GHQ-12) were provided for ratings. The result showed that WEMWBS received the highest rating among these three measures, with a median score of 7.5 for both initial and final ratings. The GHQ-12 received the lowest ratings of its appropriateness, with a score of 3 and a score of 4 for the initial rating and final rating respectively. Participants also claimed that they

preferred WEMWBS in the sense that the nature of items is related to positive spectrum of mental health, which is a favourable feature as they felt upset when answering questions about difficulties related to mental illness or negative mental health.

WEMWBS has been widely applied in different sectors of the economy such as health, business and education (Shah & Stewart-Brown, 2017). It was developed with funding support provided by the Scottish Government for the development of mental health policies. It is recommended and used by policy makers in England, Wales and Scotland (Diana Bardsley *et al.*, 2017; Parkinson, 2007; Scottish Government, 2018), as a national indicator for monitoring national mental well-being in which an increase in WEMWBS score becomes one of the health targets for the Scottish Government and in England. In terms of statistical data, WEMWBS/SWEMWBS is becoming more popular based on the annual trends of usage (Shah *et al.*, 2017b). Between 2006 and 2012, usage of the WEMWBS/SWEMWBS remained at a level below 200 per year. However, the number of registrations has been increasing at an astonishing rate since 2012, with 1328 registrations in 2016. In 2019, the number of registrations reached 200 per month. An increasing trend of the number of WEMWBS/SWEMWBS publications since the first published article in 2008 is also revealed, with a total number of 215 publications in 2016. Among all the WEMWBS/SWEMWBS interventional studies, mental health education and psychological therapies are the two most common types of intervention evaluated. Around 350 licences are now issued monthly for the use of this scale across 50 countries (Stewart-Brown, 2021). The scale has also been translated into 36 languages already, acting as an international tool for measuring mental well-being.

In addition, WEMWBS/SWEMWBS focuses on the assessment of positive mental health, which is suitable for capturing the benefits of health promotion interventions related to mental well-being in economic evaluations. As stated before, although there are other mental well-being instruments, WEMWBS/SWEMWBS has a unique and broader coverage of eudemonic well-being. In terms of face validity, WEMWBS/SWEMWBS is the closest match to what it claims to measure, which is generic mental well-being. As discussed previously, compared to WEMWBS/SWEMWBS, WHO-5 has a relatively limited coverage of mental well-being. SHES is related to school environment and MHC-SF is not much used as an outcome for mental health promotion. All these elements contribute as building blocks to the strength of WEMWBS/SWEMWBS.

Furthermore, the need to value WEMWBS/SWEMWBS states in cost-benefit analysis has been recognised by members the Housing Associations' Charitable Trust. They are currently developing the social values for WEMWBS/SWEMWBS, based on its relationship with life satisfaction (Daniel Fujiwara *et al.*, 2017; Lizzie Trotter *et al.*, 2017). The unique popularity of WEMWBS/SWEMWBS strengthens the motivation to develop a preference-based tariff for the cost-utility analysis.

There has been an increase in the number of studies that aim to evaluate the psychometric properties of WEMWBS and SWEMWBS in different population samples. In the following section, the published empirical and theoretical evidence will be reviewed on the merits and acceptability of WEMWBS and SWEMWBS across different population groups, so as to further highlight the popularity of the WEMWBS/SWEMWBS for the measurement of mental well-being.

2.6.1. Evidence to support the use of WEMWBS and/or SWEMEBS in measuring mental well-being

WEMWBS and SWEMWBS have been widely validated in different population groups across the world. Overall, they were shown to be responsive to mental health interventions, lack of floor and ceiling effects, uni-dimensional, high internal consistency, good construct validity, good test-retest reliability and high face validity, etc. A selected number of relevant publications is summarised in Appendices 4-6.

2.7. Conclusion

This chapter reviewed the theoretical framework of preference-based approaches and compared existing preference-based instruments and non-preference-based mental well-being instruments. WEMWBS/SWEMWBS was identified as the best candidate for preference elicitation in terms of its popularity, wide coverage of eudemonic well-being, and robust psychometric properties. The valuation set generated by WEMWBS/SWEMWBS would be sensitive in capturing the mental well-being benefits within an intervention.

Chapter 3: An overview of research methods for constructing the valuation set

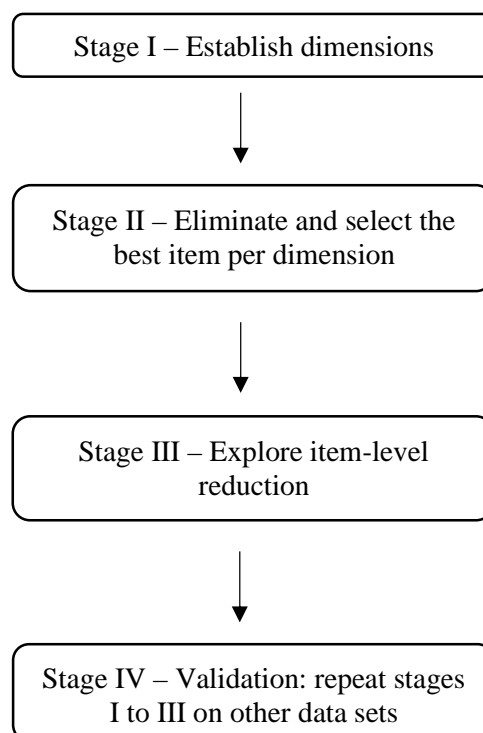
3.1. Introduction

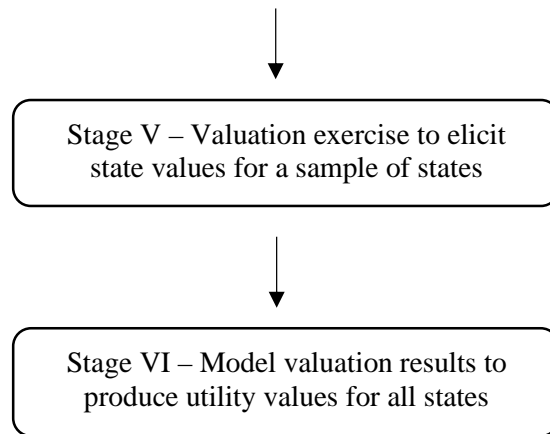
In the previous chapter, justification was provided for developing a preference-based tariff for the WEMWBS/SWEMWBS in order to allow the estimation of MWALYs for use in economic evaluation of mental well-being interventions. This chapter aims to describe the methods and strategies of the tariff development process.

3.2. The six stages of the construction of a preference-based tariff for the WEMWBS/SWEMWBS

The criteria or the required procedure for the establishment of a preference-based tariff was informed by the six-stage method proposed within a recent Health Technology Assessment report (Brazier *et al.*, 2012). An overview of the six stages is shown in Figure 1 and each of these stages will be discussed comprehensively. It is noted that Stage I to Stage IV have been completed by previous research and a summary review of these four stages will be provided.

Figure 1: The six stages of the development of a preference-based instrument





Source: Reproduced from Brazier *et al.* (2012)

3.2.1. Stage I: Establish dimensions

Conceptually, the idea of establishing structurally independent dimensions of a well-being state classification system is required for the valuation and modelling stages during the process of preference elicitation. The implementation of an exploratory or confirmatory factor analysis is the main approach for interpreting the dimension structure of a measure. Based on the evidence of psychometric validation, factor analyses have been conducted in a number of WEMWBS studies with datasets across different countries e.g. (Bass *et al.*, 2016; Clarke *et al.*, 2011; Hoffman *et al.*, 2019; Koushede *et al.*, 2019; Taggart *et al.*, 2013; Tennant *et al.*, 2007a; Waqas *et al.*, 2015). While a one factor model was confirmed by these studies in general to support the uni-dimensionality of the WEMWBS in measuring mental well-being, multidimensionality was also detected for some of the items in WEMWBS e.g. (Bass *et al.*, 2016; Stewart-Brown *et al.*, 2009), providing one of the driving forces for exploring the reduction of potential items in the later stages.

3.2.2. Stage II: Eliminate and select the best items per dimension

After the establishment of the dimensionality of the instrument in stage I, it is important to assess the fitness of items to each independent dimension and ensure a minimum number of selected items are representative of the underlying dimension. While the WEMWBS incorporates a wide coverage of information related to aspects of hedonic and eudemonic well-being, it could theoretically generate 6,103,515,625 (5^{14}) well-being states, the valuation of which is deemed to be impractical using existing valuation techniques as the valuation cost or respondent burden would be unaffordable given the huge number of well-being states. Because of this, an appropriate version of the

WEMWBS with a manageable number of well-being states and favourable measurement properties is required.

It is recognised below that there are two published studies that discuss the validity of WEMWBS with attempts to follow closely from stages II to IV for achieving the goal of establishing a feasible instrument amenable to valuation through the use of the Rasch analysis. Rasch analysis is a technique that converts qualitative (categorical) responses to a continuous (unmeasured) latent scale using a logit model (Andrich, 1981; Tesio, 2003), so as to eliminate items within a dimension with unsatisfactory representation of the underlying latent scale. The elimination criteria are based on identifying items that are not suitable for item-level ordering, investigating items split for differential item functioning in which the characteristics of respondents such as age, sex and educational level contribute to the difference in responses to a particular item, and judging the Rasch model goodness-of-fit statistics. For the selection of the best items per dimension, several criteria can be assessed including feasibility, internal consistency, and distribution of responses, etc.

The first relevant study by Stewart-Brown *et al.* (2009) analysed the internal construct validity of the WEMWBS using a number of fit statistics that estimated the level of fitness of WEMWBS to the Rasch measurement model. Data were collected from Wave 12 of the Health Education Population Survey, based upon a representative sample of the 779 Scottish adults. Additionally, the second study by Bartram *et al.* (2013) applied Rasch analysis to further validate the WEMWBS in the UK veterinary profession, with the aid of data from two independent cross-sectional surveys: a postal questionnaire survey and the Royal College of Veterinary Surgeons Survey of the Profession 2010.

The evidence obtained from these two studies will be summarised in each stage covering stage II to stage IV outlined by Brazier *et al.* (2012) for the development process of a preference-based instrument, as shown in Appendix 7.

The conclusion of this stage is that both studies agreed unanimously with the reduction of the 14 items in WEMWBS to the 7 items (Items 1, 2, 3, 6, 7, 9 and 11) in SWEMWBS.

3.2.3. Stage III: Explore item-level reduction

It is crucial to ensure that the number of response levels for each item is appropriate and not overwhelming in the sense that respondents can be able to differentiate the meaning of different response choices. To ensure this, the distribution of response frequencies or threshold probability curves should be investigated so that item levels that are closely attached to each other, or with low response frequencies, are potential candidates for level collapsing.

The main finding of this stage is that both studies supported the five-level version of the SWEMWBS, without the need to collapse the number of levels in the descriptive system.

3.2.4. Stage IV: Validation: repeat stages I to III on other data sets

After the generation of a reduced classification system, it is important to confirm the validity of the reduced items before moving on to the valuation task. This can be informed by repeating stages I to III within an alternative sample of the same dataset, or a sample from an alternative dataset.

The main finding of this stage is that both studies confirmed the validity of SWEMWBS by testing in alternative sample datasets.

A shorter version of WEMWBS has been developed based on a comprehensive Rasch analysis described by these two studies, without dropping items that have a strong independent impact on the total level of well-being within the scale. It is apparent that the SWEMWBS, which is presented in Appendix 8, could be a suitable instrument for valuation in the U.K. in the sense that it has undergone a formal Rasch analysis and the analysis results were supported by the above published U.K. empirical evidence. Also, the SWEMWBS has been further validated within different psychometric studies across service users and samples of the general population, which were discussed thoroughly in sections 2.6.1. Importantly, one of the strengths of the SWEMWBS is that it has been proved to have a strong correlation with WEMWBS, as illustrated by a spearman correlation coefficient of over 0.9 for one of the validation studies in England discussed in Appendix 6 (Ng Fat *et al.*, 2017). This implies that this reduced classification system demonstrated an extensive coverage of the construct of mental well-being covered by the full 14-item version, supporting the item efficiency. However, given the fact that a 7-item scale with 5 levels per item could potentially generate 78,125 (5^7) well-being states, it is

worth noting that an effective valuation strategy can be implemented at a later stage to minimise the valuation effort even though the number of mental well-being states have been significantly reduced relative to the 14-item version.

3.2.5. Stages V and VI: Valuation exercise to elicit state values for a sample of states & model valuation results to produce utility values for all states

In order to provide a comprehensive derivation of the most appropriate valuation strategy for the SWEMWBS, this section will be further structured into different parts. Firstly, the description of different widely used direct and indirect preference elicitation techniques of health states will be reviewed (Arnold *et al.*, 2009; Brazier *et al.*, 1999; Brazier *et al.*, 2017; Green *et al.*, 2000). The selection of valuation techniques for SWEMWBS will be informed by the elimination of unsuitable or relatively inferior techniques, in terms of their strengths and limitations in the application of previous health state valuation studies. A justification for the most appropriate administrative technology during the interview procedure of the preference elicitation task will also be provided. After that, alternative experimental designs for the selection of SWEMWBS states for valuation will be described. Finally, one qualitative and one quantitative piloting phases for testing the validity of proposed valuation protocol will be outlined.

3.2.5.1. Identification of the appropriate valuation techniques for mental well-being states

Different valuation techniques could theoretically yield different valuation results. In general, existing valuation techniques can be grouped into two categories: direct valuation methods and indirect valuation methods.

3.2.5.1.1. *Direct valuation methods*

Direct valuation methods are used to measure the strength of preference for a health state by assigning a weight for a particular health state directly onto a scale anchored at 0 (representing a death state) and 1 (representing a full health state), in which negative values represent states considered worse than dead. A respondent completes a valuation task concerning their current own health. Five direct valuation techniques, which have been commonly applied to health preference measurement, are identified: visual analogue scale (VAS), magnitude estimation (ME), standard gamble (SG), time trade-off (TTO) and person trade-off (PTO). A review of these techniques is provided in Appendix 9.

3.2.5.1.2. Indirect valuation methods

Instead of deriving a value for a particular state directly based on the respondent's valuation, indirect valuation methods derive valuations indirectly in multiple steps in the sense that the value is not solely calculated by a single formula based on a respondent's valuation. The values obtained for hypothetical health states can be modelled (or otherwise analysed) to enable utilities to be generated for every health state defined by the classification system. This can include mapping preferences onto a utility scale indirectly through the use of an existing generic HRQoL questionnaire. They can also include ordinal or choice-stated methods in which health states are ranked and selected to reflect the most preferred health state before a value is elicited. Three common indirect valuation techniques have been identified: mapping, discrete choice experiments (DCEs) and best-worst scaling (BWS). A review of these techniques is provided in Appendix 10.

3.2.5.1.3. Justification of the valuation techniques for SWEMWBS

This section aims to provide justification regarding the most appropriate methodology to be applied to the valuation of the SWEMWBS's mental well-being states. As this was the first attempt to derive a preference-based tariff for SWEMWBS states, there was no published protocol or template available. The selected well-being valuation protocol was judged against a trade-off between the strengths and limitations of valuation techniques and the lessons learnt from the previous experiences of health state valuation studies.

3.2.5.1.3.1. Mapping

To begin with, it seems attractive at first glance that mapping between SWEMWBS and an existing generic preference-based instrument may be an efficient valuation method as it requires only a selection of an appropriate preference-based instrument for econometric analysis, without spending time on the collection of primary data regarding preference values elicited from the general population. However, in order to ensure the robustness of the statistical result, mapping also requires a sufficiently large dataset with both the SWEMWBS and an existing generic preference-based instrument. Moreover, as mentioned above, the effectiveness of the mapping result depends on the extent of construct overlap between the two instruments. In this sense, it is difficult to identify a suitable generic preference-based MAU instruments from those discussed in the previous chapter for SWEMWBS to map onto because these instruments are characterised by their

arguably significant focus on dimensions of physical health. While SWEMWBS covers broad aspect of well-being, especially mental well-being, a less representative utility value would probably be generated under the mapping approach as there are distinctions between MAU and mental well-being measures. There was a first attempt by Johnson *et al.* (2016) to map the WEMWBS onto the EQ-5D-3L, but it was not surprising to find that the mappings were limited as respondents who scored the maximum score for the EQ-5D-3L actually reported a range of mental well-being levels for the WEMWBS. It might be worth analysing the possibility of mapping between SWEMWBS and the modified EQ-5D-5L or other MAU instruments or even the preference-based capability measures described in the previous chapter, but it is likely to be more beneficial or sensible to collect primary preference data for mental well-being states due to the absence of preference-based values for mental well-being measures at this time.

Additionally, previous studies regarding the development of population value sets for specific MAU instruments also adopted different direct or indirect preference elicitation methods. Table 3 below gives an overview of the valuation techniques apart from mapping used by the preference elicitation studies for the most recent versions of the preference-based MAU instruments and capability measures discussed in sections 2.3.1.1 and 2.3.1.2. As the valuation techniques used to develop population values might vary across different countries, the valuation studies for the same MAU instrument across each country or region are summarised independently.

Table 3: A summary of the valuation strategies adopted by the MAU instruments and capability measures

| MAU instruments/ preference-based capability measures | Reference | Year of data collection | Country or region targeted by the valuation study | Specified valuation technology | Valuation techniques |
|--|----------------------------------|--|--|--|---|
| QWB-SA | Seiber <i>et al.</i> (2008) | 1990 | United States | NA | VAS |
| EQ-5D-5L | Al Shabasy <i>et al.</i> (2021) | 2019-2020 | Egypt | EQ-VT 2.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Andrade <i>et al.</i> (2020) | 2018 | France | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Augustovski <i>et al.</i> (2020) | 2018-2019 | Peru | EQ-VT 2.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Augustovski <i>et al.</i> (2016) | 2013 | Uruguay | EQ-VT 1.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Finch <i>et al.</i> (2022) | 2020-2021 | Italy | EQ-VT administered via videoconferencing | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Bouckaert <i>et al.</i> (2021) | 2018-2020 | Belgium | EQ-VT 2.1 | C-TTO (Conventional |

| | | | | | |
|--|--|---------------|----------|-----------|---|
| | | | | | TTO + lead-time TTO) and DCE |
| | Devlin <i>et al.</i> (2018) | 2012 | England | EQ-VT 1.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Ferreira <i>et al.</i> (2019) | 2015- 2016 | Portugal | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Golicki <i>et al.</i> (2019) | 2016 | Poland | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Gutierrez- Delgado <i>et al.</i> (2021) | 2019 | Mexico | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Hobbins <i>et al.</i> (2018) | 2015- 2016 | Ireland | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Jensen <i>et al.</i> (2021) | 2018- 2019 | Denmark | EQ-VT 2.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Kim <i>et al.</i> (2016) | 2013 | Korea | EQ-VT 1.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Lin <i>et al.</i> (2018) | 2017 | Taiwan | EQ-VT 2.0 | C-TTO (Conventional |

| | | | | | |
|--|---------------------------------------|---------------|---------------|-----------|---|
| | | | | | TTO + lead-time TTO) and DCE |
| | Ludwig <i>et al.</i> (2018) | 2015 | Germany | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Luo <i>et al.</i> (2017) | 2012 | China | EQ-VT 1.0 | C-TTO |
| | Mai <i>et al.</i> (2020) | 2017 | Vietnam | EQ-VT 2.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Pattanaphesaj <i>et al.</i> (2018) | 2013- 2014 | Thailand | EQ-VT 1.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Pickard <i>et al.</i> (2019) | 2017 | United States | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | <i>Publication expected</i> | 2019- 2020 | India | EQ-VT 2.1 | NA |
| | Purba <i>et al.</i> (2017) | 2015 | Indonesia | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Ramos-Goni <i>et al.</i> (2018) | 2012 | Spain | EQ-VT 1.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |

| | | | | | |
|-------------|-----------------------------------|---------------|-------------|-----------------------|---|
| | Rencz <i>et al.</i> (2020) | 2018- 2019 | Hungary | EQ-VT 2.1 | C-TTO |
| | Shafie <i>et al.</i> (2018) | 2016 | Malaysia | EQ-VT 2.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Shiroiwa <i>et al.</i> (2016) | 2013 | Japan | EQ-VT 1.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Versteegh <i>et al.</i> (2016) | 2012 | Netherlands | EQ-VT 1.0 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Welie <i>et al.</i> (2020) | 2019 | Ethiopia | EQ-VT 2.1/ EQ- PVT | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Wong <i>et al.</i> (2018) | 2014 | Hong Kong | EQ-VT 1.1 | C-TTO (Conventional TTO + lead-time TTO) and DCE |
| | Xie <i>et al.</i> (2016) | 2012 | Canada | EQ-VT 1.0 | C-TTO (Conventional TTO + lead-time TTO) and traditional (or conventional) TTO |
| | Yang <i>et al.</i> (2021) | 2021 | Uganda | EQ-PVT Lite | C-TTO |
| HUI3 | Le Galès <i>et al.</i> (2002) | 1999 | France | NA | VAS and SG |

| | | | | | |
|----------------|--|----------------|--------------------|----|--|
| AQoL-8D | Richardson <i>et al.</i> (2014) | NA | Australia | NA | VAS and time trade-off |
| SF-6D | Brazier and Roberts (2004); Brazier <i>et al.</i> (2002) | NA | United Kingdom | NA | SG |
| | Brazier <i>et al.</i> (2009) | NA | Japan | NA | SG |
| | Cruz <i>et al.</i> (2011) | 2007 - 2008 | Southern Brazil | NA | SG |
| | Ferreira <i>et al.</i> (2010) | 2006 | Portugal | NA | SG |
| | Lam <i>et al.</i> (2008) | 2002 - 2003 | Hong Kong | NA | SG |
| | Mendez <i>et al.</i> (2011) | 2007 | Spain | NA | Inverse probability weighting |
| | Norman <i>et al.</i> (2014) | NA | Australia | NA | DCE |
| | Perpinan <i>et al.</i> (2012) | NA | Spain | NA | The probability lottery equivalent method, VAS and SG |
| 15D | Sintonen (1995); Sintonen (2001) | NA | Finland | NA | Rating scale (VAS), ME |
| ReQoL | Keetharuth <i>et al.</i> (2021) | NA | United Kingdom | NA | C-TTO (Conventional |

| | | | | | |
|-----------------|--------------------------------|----------------|-------------------|----|------------------------------------|
| | | | | | TTO + lead-time TTO) with props |
| ICECAP-A | Flynn <i>et al.</i> (2015) | NA | United Kingdom | NA | BWS |
| ICECAP-O | Coast <i>et al.</i> (2008a) | 2005 - 2006 | United Kingdom | NA | BWS |
| ASCOT | Netten <i>et al.</i> (2012) | From 2009 | United Kingdom | NA | TTO, BWS |

AQoL-8D indicates Assessment of Quality of Life - 8D; ASCOT, Adult Social Care Outcomes Toolkit; BWS, best-worst scaling; C-TTO, composite time trade-off; 15D, 15-dimensional instrument; DCE, discrete choice experiment; EQ-5D-5L, EuroQol five dimension 5-level version; EQ-PVT, EuroQol Portable Valuation Technology; EQ-VT, EuroQol Valuation Technology; HUI3, Health Utilities Index Mark 3; ICECAP-A, Investigating Choice Experiments Capability Measure for Adults; ICECAP-O, Investigating Choice Experiments Capability Measure for Older people; ME, magnitude estimation; NA, not applicable; QWB-SA, The Quality of Well-Being Scale Self-Administered; ReQoL, Recovering Quality of Life; SF-6D, Short-Form Six-Dimension; SG, standard gamble; TTO, time trade-off; VAS, visual analogue scale.

3.2.5.1.3.2. Visual analogue scale (VAS)

Among the most widely used direct valuation techniques described in the previous section or adopted by the most recent version of the MAU and capability measures, it is decided not to use VAS for valuing mental well-being states as it is arguably regarded as a non-preference-based valuation technique (Tolley, 2009). It is not a choice-based method and the valuation result is only weakly based on utility theory, if at all. In other words, the application of this technique generates values rather than utilities. Although some attempts have been shown to develop adjusted VAS ratings for the placement of a 0 (dead) to 1 (full health) scale (e.g. Williams (2005)), this kind of transformation is argued as a mere practice of mathematical rescaling, without sufficient economic theoretical support. In this context, VAS has typically been used as a warm-up task for respondents before formal valuation exercises. Other valuation techniques with a more solid theoretical framework should therefore be considered instead.

3.2.5.1.3.3. Magnitude estimation (ME)

ME is not considered for the valuation of mental well-being states because of its lack of economic theoretical support and concerns about the underlying assumptions behind this technique. It is not a choice-based task and therefore provides no economic foundation of consumer theory in decision making. Also, the assumption of ratio scaling is questionable and it is ambiguous or “obscure” in the interpretation of the ratio of undesirability between mental well-being states, due to the presence of subjective responses and the absence of a universally consistent scale for the valuation of well-being states (Richardson, 1994). It is expected that these issues are applicable to the valuation of mental well-being states, generating potential uncertainty and hurdles in the interpretation of the valuation results.

3.2.5.1.3.4. Person trade-off (PTO)

The PTO seems amenable to the valuation of societal preferences as this technique is based on deriving values for informing social welfare. However, evidence showed that there was a huge cost of implementing this technique in practice in terms of requiring a large groups of subjects to minimise the measurement error, and framing effects in the decision of starting point within the PTO exercise, etc (Nord, 1995). Also, due to the question framing of valuing choices of states that involve third parties or other people, I would argued that this is not the best valuation technique for SWEMWBS states because of the emphasis on

the self-reported nature of the SWEMWBS. The 7 items in the SWEMWBS are related to personal feelings and thoughts so that the items are all phrased as “I’ve been” at the beginning of the statement. In this sense, it might be better to explore valuation techniques that are directly related to the valuation of an individual’s own well-being states. In other words, it is better to explore a valuation technique that requires the imagination of respondents being in a mental well-being state on their own, rather than imagining the others being in that mental well-being state.

3.2.5.1.3.5. Standard gamble (SG)

SG arguably has the strongest economics foundation as it was developed based on expected utility theory. However, due to the nature of incorporating probabilities within the choice task, it can be difficult for the general public lacking learning experiences on probabilities to interpret probability theory correctly. Attempts have been made to develop a version of the SG in which respondents were presented with a probability wheel to simplify the task through the use of this visual aid (Torrance, 1976; Torrance, 1986). The “ping-pong” method was developed to help respondents reach the indifference point between two options. Jones-Lee *et al.* (1993) presented the respondents with values of chances of success and allowed them to indicate the values of choosing and rejecting treatments, with a view to deriving the value which is most difficult for them to choose across treatments. However, even if the respondent burden of interpreting probabilities is minimised, there are still some concerns surrounding the theoretical framework underpinning the SG. For example, the axioms of consumer theory can be violated as respondents might not have constant proportional risk posture during the valuation task. Also, again, the utility rescaling and bounding of states worse than dead to -1 lacks theoretical support and the result might not be highly representative.

3.2.5.1.3.6. Time trade-off (TTO)

Alternatively, TTO is simpler to adopt in practice because it avoids the difficulties in explaining probabilities to respondents. However, the underpinning economic foundation is weakened without the incorporation of uncertainty in the decision making process (Mehrez & Gafni, 1991). Also, it is subject to the level of duration effect during the valuation task, in which respondents are not necessarily willing to trade off a constant proportion of the remaining years to gain an improved quality of life (Bleichrodt *et al.*,

2003; Dolan & Stalmeier, 2003; Spencer, 2003). The trade-off process becomes more impractical given a short duration of life due to evidence of respondents' lexicographic preferences. In addition, the valuation of years of life is also affected by the direction of the rate of time preference of respondents. Moreover, similar to the SG, TTO can theoretically have unbounded negative utility values for states worse than death, which is problematic because of the imbalance between positive and negative values.

Given the fact that both SG and TTO are subject to different kinds of limitations, a myriad of evidence has suggested that TTO performs slightly better than the SG in valuation tasks (Reed *et al.*, 1993; Richardson, 1994; Vanderdonk *et al.*, 1995), although there is no definitive conclusion around a preferred valuation technique. Furthermore, for the derivation of the QALY, the TTO approach has been recommended by NICE as a choice-based method to value health states, in order to retain methodological consistency with the methods used to derive underpinning preference weights for the EQ-5D (The National Institute for Health and Care Excellence, 2013). Over and above that, most importantly, it is encouraging that the lead-time TTO and lag-time TTO have been developed to tackle the problem of the conventional TTO so that the utility values can be anchored between -1 and 1 with the advantage of a stronger theoretical support than forced rescaling. The problem of utility exaggeration can also be resolved by the new variants of the conventional TTO as the conventional version has suffered from the problem of using completely different trade-off tasks between the valuation of states worse than death, and states better than death (Devlin *et al.*, 2011). The focusing effect can be introduced to respondents and the validity of aggregating positive and negative utility values generated by different utility functions can then be questionable due to inconsistent calculation methodologies. Specifically, the time period x is introduced to the numerator of the formula used to calculate the value of states better than death, whereas it is introduced in both the numerator and denominator for the case of states worse than death. The interval properties are indefensible in the sense that the change in negative value cannot be compared to the change in positive value with the same numerical distance (e.g. the change from -0.5 to -0.4 is no longer identical to the change from 0.5 to 0.6).

3.2.5.1.3.6.1. Lead-time versus lag-time TTO

Motivated by the recognition and theoretical improvement of the conventional TTO method, it was proposed to include a composite time trade-off (C-TTO) for the valuation

of mental well-being states. The C-TTO comprises the conventional TTO for valuing mental well-being states better than death and either a lead-time TTO or a lag-time TTO for valuing mental well-being states worse than death. The reasons for not using lead-time TTO or lag-time TTO for mental well-being state considered both better than and worse than death were informed by the published evidence about its practicality for valuation studies. For example, the general conclusion from studies that applied lead-time TTO for valuation was that respondents were found to trade off the lead time even if the states were not obviously severe (Augustovski *et al.*, 2013; Devlin *et al.*, 2013; Oppe *et al.*, 2014), reflecting their difficulty in recognising the fact that trading of lead time implied valuing a state considered worse than death. The problem of obtaining a negative value even for states considered better than death can be avoided by reserving the use of conventional TTO for valuing states considered better than death, while adopting lead-time or lag-time TTO for valuing states considered worse than death. Although this method of valuation accentuates the problem of using two different elicitation tasks for valuation of states considered better than or worse than death in comparison to the conventional approach, C-TTO is regarded as preferable in the sense that a larger weighting proportion towards negative utilities can be avoided.

Regarding selection of lead-time and lag-time TTO for the valuation of mental well-being state considered worse than death, considerations in terms of the application performance and impact of the usage on valuation outcomes were taken into account as respondents' preferences for the importance of mental well-being at the end or earlier stages of life might affect the values generated by the two variants (Tilling *et al.*, 2008). The proposed decision could be informed by the published evidence.

Augustovski *et al.* (2013) investigated the influence of adopting these two variants of TTO for the valuation of EQ-5D-5L health states across a sample of the Argentinian population, as a part of the EQ-5D-5L pilot study. Lead-time and lag-time TTO were compared using mean observed values, valuation completion time, responses to follow-up and feedback questions, number of steps to arrive an indifference point for the TTO valuation, etc. The *t*-test results showed that the differences in values generated by the two variants were statistically insignificant in general. In addition, Versteegh *et al.* (2013) compared the performance of the conventional TTO with duration of 10 years, lead-time and lag-time TTO with durations of 15 and 20 years for the EQ-5D-5L health states across a sample of

the Dutch population through the administration of an online experiment. It was realised that the lead-time TTO produced higher values than lag-time TTO, and the differences increased over longer time frames. The mean absolute difference in value was the smallest between conventional TTO and the lag-time TTO over a duration of 20 years. Also, the mean variance of the general values for the health states obtained from the lag-time TTO was statistically and significantly higher than the lead-time TTO, under the durations of both 15 and 20 years. Furthermore, a study also examined the comparison of these two variants of TTO in terms of different ratios of lead or lag time to duration of EQ-5D health states (Devlin *et al.*, 2013). The justifications regarding the merits and limitations of both methods were mixed. For example, the lead-time TTO was regarded as more realistic than the lag-time TTO for the valuation of severe health states. It is more sensible to interpret a poorer health state after exhibiting a period of full health, rather than returning to perfect health after suffering from extremely unfavourable health states such as not being able to walk, as some health states can be permanent. The lag-time TTO was discovered to reduce the non-trading effect for mild states. However, it also suffered from more inconsistencies between the signs of utility values and respondents' judgment of the health state considered better or worse than dead, relative to the lead-time TTO. Discounting had a relatively larger effect on lag-time TTO values in which the differences between discounted and undiscounted values were larger. The authors emphasised the difficulty in the identification of the best TTO variant and judgments regarding the importance of different characteristics of the valuation data generated by these two variants are required.

Although there is no existing evidence commenting on the preferred or recommended variant of the TTO approach, lead-time TTO was used for valuing well-being states worse than death as the lead-time TTO has been successfully applied in some published valuation studies including those recently published EQ-5D-5L valuation studies in Table 3 and published evidence regarding its validity and feasibility has begun to arise (Janssen *et al.*, 2013). Conversely, there is still a lack of supporting evidence around the validity of the lag-time TTO in the application of valuation studies, which constrains confidence about the use of this variant of TTO until it is more mature or developed in practice.

3.2.5.1.3.6.2. Time horizon and duration of well-being state

Concerning the time horizon of the C-TTO task, ten years was set for the conventional TTO valuation of better-than-dead mental well-being states and twenty years was set for

the lead-time TTO valuation of the worse-than-dead mental well-being states, as suggested by the EuroQol group for the TTO valuation of EQ-5D-5L states (Oppe *et al.*, 2016). For the ratio of lead time to duration of mental well-being state, a ratio of 1:1 (10 years for both the lead time and duration) adopted by the EuroQol group was followed so as to avoid imbalanced weighting of positive and negative utility values. The visual representation of the C-TTO task is presented in Figure 2. The C-TTO task began by allowing respondents to choose or indicate indifference between life A (10 years in full mental well-being) and life B (a mental well-being state which is worse than full mental well-being), as illustrated in Figure 2a. If respondents exhausted all the 10 years of full health and still failed to reach the indifference point, the mental well-being state was regarded as worse than death and a lead-time of 10 years was added to allow extra trade-off of full mental well-being, as illustrated in Figure 2b.

Figure 2: The C-TTO layout

Figure 2a: Conventional TTO for the valuation of mental well-being states considered better than dead

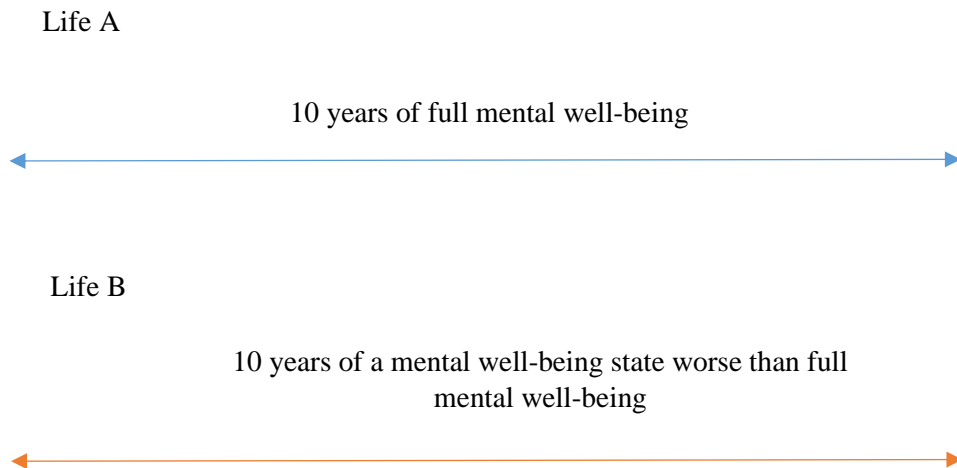
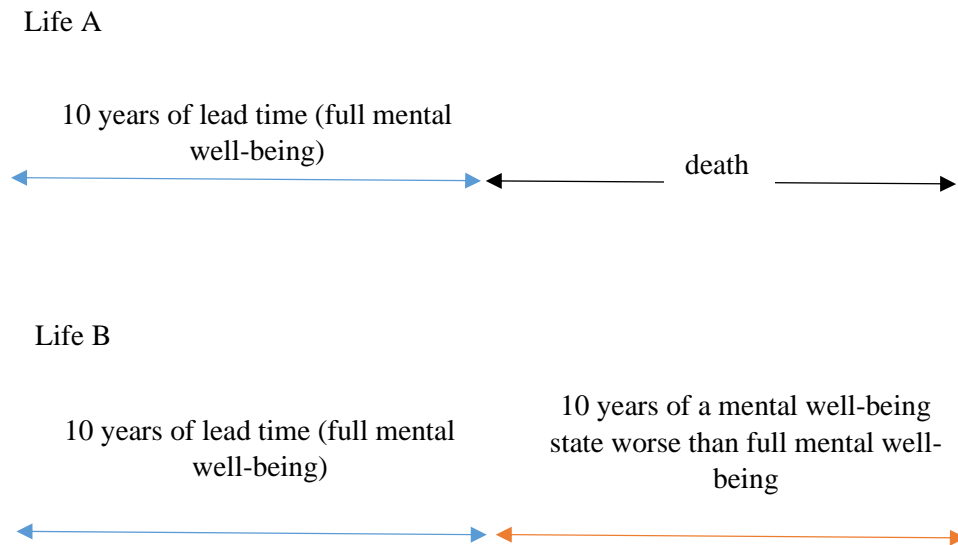


Figure 2b: Lead-time TTO for the valuation of mental well-being states considered worse than dead



Although there have been some attempts to incorporate variants of a higher ratio of lead-time to mental well-being state due to worries over the exhaustion of lead time and the existence of a potential large negative utility values (Devlin *et al.*, 2013), it was unnecessary for the valuation of the SWEMWBS because of its positive framing of statements. Possibly, even the lowest mental well-being state could be rarely valued as worse than death by respondents, let alone generated a large negative value in this sense. The utility value of respondents who exhausted all lead time in the valuation of mental well-being states would be censored at -1 as only small number of cases with values falling below -1 was expected. However, it could be argued that the suffering attributable to extremely low mental well-being was a source of even greater distress than that attributable to severe physical health. Because of this, if a respondent was regarded as failing to achieve an indifference point after the exhaustion of lead time, an extra question regarding the length of time in full mental well-being required to compensate or choose the option with the lower mental well-being state was asked. The decision surrounding whether the ratio of 1:1 is suitable for the C-TTO task was informed by the piloting stage.

3.2.5.1.3.6.3. Iteration algorithm

Consistent with the protocol adopted by the EuroQol Group, a combination of bisection and titration methods were chosen (Oppe *et al.*, 2016). The bisection approach was applied to the first three steps of the C-TTO task, followed by upward or downward titration with an increment of either 1 year or 6 months. When preference reversal occurred in any steps of the task, there was a correction period of 6 months. The ping-pong approach was not considered for the iteration as previous evidence showed that there was a potential increase in respondent burden in terms of a longer completion time (Lenert *et al.*, 1998). Moreover, a higher variability of utility was observed due to the nature of this ping-pong procedure. A larger sample size than titration procedure was also required to detect a utility difference between health states. These findings were highly unfavourable for this project as efficiency was a key concern during the huge and complicated procedure of primary data collection at later stages.

3.2.5.1.3.7. Discrete choice experiments (DCEs) and Best-worst scaling (BWS)

In addition to the above direct valuation techniques, there is increasing interest in the implementation of indirect valuation techniques. DCEs are one of the preference elicitation techniques (Ryan & Gerard, 2003) that have been widely used in the area of marketing research and their application in healthcare valuation research is increasingly evolved (Bahrapour *et al.*, 2020; Mulhern *et al.*, 2019). Evidence showed that DCEs are simpler for respondents to complete as they are less cognitively challenging for respondents and can reduce the time needed to compare scenarios and derive valuations (Bansback *et al.*, 2012). BWS is another choice-based technique with an increasing interest for application in healthcare valuation studies. Due to the unique advantages and perspectives of both techniques over the traditional valuation techniques adopted in the past decades, they are worth pursuing in the valuation study of the SWEMWBS.

Among the choices of DCEs and BWS, it was decided to adopt DCEs for the valuation of mental well-being states. The main driving force behind this decision was that the DCE is also one of the valuation techniques adopted by the preference elicitation studies for the EQ-5D-5L (Devlin & Krabbe, 2013). With a view to aid comparability between QALYs and MWALYs, it is worth closely following the most recent valuation protocol for the EQ-5D-5L, so as to explore potential differences and implications of the preference-based

tariff derived from the EQ-5D-5L MAU instrument and the SWEMWBS well-being instrument. Also, it is argued that the BWS is based on a weaker economic foundation as there is no trade-off task included in the mere indication of the best and worst choices (Coast *et al.*, 2008a). The concept of opportunity cost is therefore absent in the decision made by the within-profile choices. Moreover, although there is evidence implying the similarity of the preference weights generated using social care data by the two methods (Potoglou *et al.*, 2011), a number of comparative studies in the area of health economics support the relative superiority of the DCE. For example, Krucien *et al.* (2017) aimed to compare the validity of the DCE and the BWS in terms of the valuation of glaucoma-related health states. Although both methods were similar in the performance of preference completeness, the BWS performed relatively poorer in the measurement properties of stability, monotonicity and continuity. A higher proportion of respondents failed in the stability test of the BWS (BWS: 24% v.s. DCE: 13%), implying a poorer ability of respondents to indicate the same choices given the repeated BWS's task. In terms of monotonic preferences, it was noteworthy to discover that the proportion of respondents fully satisfied and fully failed the monotonicity test for the DCE was 73% and 2% respectively, compared to 0% and 42% respectively for the BWS responses. This implied a significant problem of indicating the dominated alternative as the more attractive or preferred. For the issue of continuity, respondents who revealed a dominant preference for a particular attribute were 23.5% and 16.6% for the BWS and the DCE methods respectively, implying a relatively higher lexicographic score for the respondents completing the BWS tasks in general. Additionally, Whitty *et al.* (2014) compared the two methods in terms of their applications in revealing Australian public preferences for the area of healthcare priority setting. The DCE task required respondents to indicate a preferred intervention to be funded among two choices with different scenarios of new technologies such as the main benefit of intervention, the ability of the intervention to provide good value for money, and how many patients are expected to benefit from the intervention, etc. For the BWS task, respondents were asked to choose the most and least important considerations when making a funding decision. While there was a weak correlation between the preferences obtained from the DCE and BWS models, the DCE task was found to be preferable to the BWS task in the context of ease of completion and response consistency. The proportion of respondents reporting the DCE tasks to be difficult or very difficult to complete and easy or very easy to complete were 22.2% and

39.4% respectively, compared to 31.9% and 28.8% respectively for the profile case BWS, driving the preference for using DCEs over BWS in more than 70% of respondents. Also, the response consistency rate for the repeated DCE tasks (75.7%) was higher than that of the BWS across the attribute levels (64.5%, 49.4% and 35.5% for most, least, and both most and least preferred attribute levels, respectively).

Based on the above analytical process, the DCE was preferred as an indirect technique for valuing the mental well-being states of SWEMWBS. A pairwise DCE with forced choice constituted the format of the choice task. Respondents were asked to choose between pairs of mental well-being states. An example of a pair of mental well-being state (2314442 v.s. 2544344) is visually provided in Table 4.

Table 4: An example of a pairwise DCE with forced choice

| Mental well-being state A (2314442) | Mental well-being state B (2544344) |
|--|---|
| <i>Rarely</i> feeling optimistic about the future | <i>Rarely</i> feeling optimistic about the future |
| <i>Some of the time</i> feeling useful | <i>All of the time</i> feeling useful |
| <i>None of the time</i> feeling relaxed | <i>Often</i> feeling relaxed |
| <i>Often</i> dealing with problems well | <i>Often</i> dealing with problems well |
| <i>Often</i> thinking clearly | <i>Some of the time</i> thinking clearly |
| <i>Often</i> feeling close to other people | <i>Often</i> feeling close to other people |
| <i>Rarely</i> able to make up my own mind about things | <i>Often</i> able to make up my own mind about things |
| Which is better, mental well-being state A or mental well-being state B? | |

3.2.5.1.4. Administrative technology for the valuation procedure

In terms of the process of conducting the valuation exercise, it was important to explore a formal and efficient procedure. It was proposed that the EuroQol Valuation Technology (EQ-VT) protocol (Oppe *et al.*, 2014), which has been used to derive different published sets of EQ-5D-5L preference-based tariffs for a number of countries or regions, was used

as the interview protocol for the valuation process of SWEMWBS's well-being states. As there are various versions of the EQ-VT, the reasons for the adoption of the EQ-VT and its specific version 2.1 will be discussed as follows.

To begin with, one of the obvious advantages for the adoption of the EQ-VT was cost reduction when the Computer-Assisted Personal Interview (CAPI) Software was utilised during the process of face-to-face interview. As responses were recorded by the computer directly, time could be saved for data entry during the process of data extraction and human error will be avoided. Also, interviewer effects due to the impact of administration mode could be reduced and a higher level of consistency in protocol compliance could be achieved. Considering the potential utilisation of a complete digital procedure during the interview process, it was decided to include a physically based interviewer even if the CAPI software was adopted during the interview, so as to maximise engagement of the respondents and allow them to get real-time immediate help whenever they encountered difficulties or queries about the valuation task (Augustovski *et al.*, 2013).

The initial version of this technology, the EQ-VT Version 1.0, was introduced in 2012 and used to derive sets of preference-based tariffs specific to Canada, England, Netherlands, China and Spain. However, concerns regarding interviewer effects on the valuation results and data quality have been raised in a number of valuation studies or review articles (e.g. Hernández-Alava *et al.* (2018); Ramos-Goni *et al.* (2017b)). Specifically, variants in interviewer behaviour and level of engagement of the respondents contributed to the clustering of values and high rates of inconsistent responses (Devlin *et al.*, 2018; Versteegh *et al.*, 2016). An alternative version named EQ-VT 1.1 was used in Japan, Korea, Uruguay, Hong Kong and Thailand as the preference elicitation technology. In this version, a quality control software was incorporated with a view to supervise the performance of interviewers and interviewees during the valuation process and their ability to follow the instructions of the protocol (Ramos-Goni *et al.*, 2017a). Interviews which were poor in quality due to the absence of explanation of the lead-time task, too short a time for interviewers to spend on explaining the C-TTO task in the wheelchair example specified by the EQ-VT, significant inconsistent valuations and an unexpected short completion time for the C-TTO tasks by the respondents were flagged up in the quality control reports. Interviewers were retrained or excluded if a constant flagging was detected. Also, three additional practice states were included before the real task to help

respondents familiarise themselves with the C-TTO tasks and the interpretation of health states. Recently, the EQ-VT Version 2.0 has been introduced and adopted to establish the tariff for France, Portugal, Poland, Mexico, Taiwan, United States, Indonesia, Germany, Ireland and Malaysia. A feedback module was added in this version, allowing respondents to raise disagreement with the rank ordering of health states implied by their responses. The most recent version at this moment is the EQ-VT 2.1 (Stolk *et al.*, 2019). A dynamic question regarding the imagination of a health state that is much better or much worse is included after the completion of the wheelchair example by the respondents. It was applied to the EQ-5D-5L preference elicitation studies in Ethiopia, Hungary, India, Vietnam, Denmark, Belgium, Peru and Egypt.

It was proposed that the version of EQ-VT 2.1 would be adopted for the valuation of SWEMWBS as it is the most up-to-date version with a strict quality control process. The validity of the valuation result could be enhanced. However, due to the nature of differences between health state and mental well-being state valuation, it was necessary to modify some of the features and contents of the EQ-VT. These will be discussed in the interview process of two piloting phases in detail. Considering the potential modifications to the EQ-VT 2.1, I was decided to call the valuation technology adopted throughout this thesis as the “adjusted EQ-VT 2.1”.

An overview of the proposed valuation strategy is presented in Table 5 below.

Table 5: Summary of the valuation strategy for the SWEMWBS

| | |
|---|--|
| Valuation technology | Adjusted EQ-VT 2.1 |
| Administration mode | CAPI with the presence of an interviewer. |
| Valuation techniques | |
| 1. <i>C-TTO</i> : Conventional TTO for the valuation of mental well-being states considered better than death and a lead-time TTO for the valuation of mental well-being state considered worse than death. | |
| ➤ Time horizon and duration of well-being state | 10 years for states considered better than death; 20 years for states considered worse than death. |

| | |
|--|---|
| ➤ Ratio of lead time to duration of mental well-being state | 1:1 (10 years for the lead time and 10 years for the duration of mental well-being state) |
| ➤ Iteration algorithm | The bisection approach will be applied to the first three steps, followed by upward or downward titration with an increment of either 1 year or 6 months. A correction period of 6 months will be applied whenever preference reversal occurs in any steps. |
| 2. <i>Pairwise DCE</i> : Different pairs of mental well-being profiles with forced choice. | |

CAPI indicates Computer-Assisted Personal Interview; C-TTO, composite time-trade off; DCE, discrete choice experiment; EQ-VT, EuroQol Valuation Technology; TTO, time trade-off.

Given the similar valuation techniques being adopted (i.e. C-TTO and DCE), the UK preference-based valuation set derived from the SWEMWBS cannot be directly compared to the current England preference-based valuation set derived from the EQ-5D-5L, due to the difference in the use of EQ-VT protocol (i.e. EQ-VT 1.0 for the valuation of EQ-5D-5L in England versus adjusted EQ-VT 2.1 for the valuation of SWEMWBS in the UK).

3.2.5.2. Selection of a sample of mental well-being states for valuation

Due to the fact that there are too many possible mental well-being states within SWEMWBS that can be valued by a single respondent in a single interview, it was necessary to select a subset of mental well-being states for valuation in order to minimise respondent burden as previous experience of health state valuation studies showed that respondents were able to value a limited number of health states. For instance, around 13 health states were affordable for the piloting of the EQ-5D-3L valuation (Dolan, 1997), and 17 DCE choice sets were manageable for the valuation of six attributes of dental services with 2 to 4 levels each (Bech *et al.*, 2011). The valuation result of this subset can be used to derive a valuation function for extrapolating the preference-based values for the remaining mental well-being states.

Since there was no published guidance on the identification of this subset, justification of the appropriate selection methodology was based on learning from previous valuation studies and their efficiency and effectiveness in application. The proposed specifications of the experimental designs for both C-TTO and DCE are discussed below.

3.2.5.2.1. *Design for the DCE*

Traditionally, orthogonal designs were applied in most of the valuation studies in health economics. However, concerns have been raised regarding the violation of orthogonality assumption under the situation of transforming categorical variables to dummy variables (Ferrini & Scarpa, 2007; Rose & Bliemer, 2004; Stolk *et al.*, 2010). In other words, orthogonal designs were typically not applicable for capturing non-linear effect for the attributes' levels.

The Bayesian efficient design was proved to be a suitable algorithm for the selection of DCE pairs in the valuation of the EQ-5D-5L (Krabbe *et al.*, 2014; Oppe *et al.*, 2014; Oppe & Van Hout, 2017). With a reference to this experimental algorithm, it was proposed to apply the D-efficient design in the selection of mental well-being states, with the minimisation of the simulated subsample D-error from the full factorial design as the main selection criterion for the DCE pairs. This design allows the relaxation of strict orthogonality assumption when modelling the dummy variables for the attributes' levels.

3.2.5.2.2. *Design for the C-TTO*

In order to derive a balanced mix of mental well-being states in which any one level of any one item had an equal chance of being combined with the levels of other items, a blocked design was used as the selection of mental well-being states for valuation.

Each participant was responsible for valuing one of the blocks. With a reference to the EQ-VT experimental design for the valuation of EQ-5D-5L health states (Oppe & Van Hout, 2017), there was a compulsory inclusion of the worst health state (55555) and one of the very mild health states (21111, 12111, 11211, 11121, 11112) within each block. Additional 8 states per block were randomly generated with the aid of the Monte Carlo simulation and the construction for level balance optimisation criterion. It is noted that the EQ-5D-5L state 11111 is the full health state, whereas the SWEMWBS state 1111111 is the lowest mental well-being state. To adapt these numbers to the design, it was decided to include the lowest mental well-being state (1111111) and one of the closer to full mental well-being states (4555555, 5455555, 5545555, 5554555, 5555455, 5555545 and 5555554) as two compulsory valuation states for each respondent within each block. The inclusion of the lowest mental well-being state was to investigate how low could the generated value be for this state, informing the econometric specification for the

regression model in extrapolating the preference-based values for the remaining mental well-being states. The inclusion of the closer to full mental well-being states was to statistically distinguish closer to full mental well-being state from full mental well-being (Devlin *et al.*, 2018). Moreover, additional states were randomly allocated to each block.

3.2.5.3. Piloting studies to validate the valuation methodology in a suitable sample

Given the lack of existing research on the issues and challenges involved in valuation of mental well-being states, it was necessary to carry out proper and extensive piloting stages to gather information on the strengths and limitations of the designed valuation strategy for preference elicitation. Specifically, the following issues were explored during this stage:

- The extent to which respondents value mental well-being states considered to be worse than death. This was important to inform whether it was appropriate to censor the negative utility value at -1. If there was a significant proportion of respondents (e.g. >10%) who fail to reach an indifference point when the lead-time is completely exhausted in the TTO task, the ratio of lead time to duration of mental well-being states at 1:1 should be reconsidered carefully. As mentioned in the proposed valuation strategy, an extra question regarding the length of time of full mental well-being required to compensate or choose the option with the low mental well-being state would be asked for those who completely exhaust the lead time. Results would reflect whether the censoring method applies to a large number of respondents with negative utility much less than -1, causing the problem of statistical invalidity to reveal general preference precisely.
- The practicality and feasibility of the adoption of the adjusted EQ-VT 2.1 during the interview process. It was important to resolve any technical issues before rolling out to the national valuation exercise.
- The number of inconsistent responses identified in the valuation exercise. Inconsistency means that a higher (lower) utility score is obtained for a mental well-being state that is logically considered as inferior (superior) than another mental well-being state (Yang *et al.*, 2017).
- The ability of the respondents to complete the valuation exercise. Feedback on issues such as wording and cognitive burden were investigated so as to update and

inform sample size calculations for the valuation study and guide prioritisation of states for inclusion in the valuation exercise.

- The total length of time for the completion of C-TTO tasks and the DCE tasks. This could inform the cost of conducting the valuation interviews.

In order to address the above concerns and carefully test the valuation protocol, the whole piloting framework was divided into two main phases, as discussed as follows:

3.2.5.3.1. Phase I (Qualitative phase): Cognitive interviews with the use of think-aloud and verbal probing techniques

The first phase of the piloting involved a comprehensive investigation of the designed valuation protocol by the use of a qualitative interviewing approach – the cognitive interview. It is defined as “*one-to-one interviews in which verbalization is used to access the thoughts and feelings, and to understand the ideas and interpretations, of respondents who are being asked to process information (Willis, 2004).*” The aim of this qualitative piloting phase was to gather the thoughts and feelings of completing the C-TTO and DCE exercises in mental well-being valuation with the use of the CAPI method. The idea was to obtain insights regarding the application of the proposed valuation protocol. Interviewees were given chances to “think aloud” the things that came into their minds during or after the completion of valuation exercises. The justification of the adoption of this interview strategy was that information regarding the cognitive process of completing the valuation tasks could be interpreted based on their verbal expressions of the thoughts about the easiness and problems of completion. Moreover, instead of interpreting the valuation results solely on the final valuation outcome, the verbal information obtained could be used to inform the understanding of the quantitative results. In order to allow interactions between interviewees and interviewer, it was decided to supplement the collection of verbal information by including follow-up probes. Using the appropriate probes for the think-aloud process could help manage the behaviour of interviewees in the sense that they could be directed back to the main discussion path in case they were distracted or diverted to irrelevant points (Willis, 1994; Willis, 2004). The efficiency of information collection could then be maximised.

The application of cognitive interview in health economics literature mainly focused on the completion of questionnaires for identifying errors and problems of questionnaire

design, with a view to refine items in questionnaires. For example, Murtagh *et al.* (2007) applied the cognitive interview with the use of both think-aloud and verbal probing techniques in palliative care research to understand the cognitive process of completing the Memorial Symptom Assessment Scale, Geriatric Depression Scale and the Palliative Care Outcomes Scale by renal patients. Content analysis was used for data analysis. Bailey *et al.* (2016) used think-aloud cognitive interviews, followed by semi-structured interviews to explore the feasibility of completing the ICECAP-Supportive Care Measure, the EQ-5D-5L and the ICECAP-A by patients receiving hospice care, close persons to the patients, and health professionals. The method of constant comparison was used to analyse themes within the interview transcripts. Also, Al-Janabi *et al.* (2013) investigated the possibility of self-reporting capabilities by adopting think-aloud with probing techniques plus semi-structured interview for respondents in the U.K. A constant comparative method was used to derive themes in thematic analysis after completing the ICEPCAP-A and the EQ-5D.

In addition, cognitive interview was also applied in the area of health state valuation. For example, Robinson *et al.* (1997) investigated the completion of the VAS and the TTO tasks for eliciting the values of the EQ-5D-3L health states. A sample of respondents who participated in the Measurement and Valuation of Health (MVH) study in the U.K. was invited to think-aloud the cognitive process of reaching the answers to the valuation tasks. The results were used to inform the comparison of the differences in VAS and TTO values among respondents with different age groups. Another study performed by Spencer (2003) also applied think-aloud techniques and probes to analyse the completion of different variants of TTO for valuing the health states of the EQ-5D-3L. The results were used to test the idea of procedural invariance, in which preferences are independent of the elicitation method. Furthermore, Janssen *et al.* (2013) empirically investigated the feasibility and validity of completing the valuation tasks in a face-to-face standardised computer assisted interview setting. Statistical properties including the mean values of the valuation results, average number of steps to finish the valuation exercise, average completion duration, and percentage of responses to a number of debriefing statements were analysed. Although these studies provided some insights into the valuation techniques, we know little about the application of health state valuation techniques into the valuation of mental well-being. This piloting phase therefore aimed to investigate the

cognitive process of completing C-TTO and DCE exercises for the valuation of the SWEMWBS to inform the optimisation of a valuation protocol.

The type of recruited respondents, the sample size involved in this phase, the strategies for the selection of SWEMWBS states for valuation, the valuation platform, the interview process of the cognitive interview, and the use of thematic analysis for the data analysis of verbal texts will be documented comprehensively in Chapter 4. The result of this phase was used to modify the proposed valuation protocol for SWEMWBS.

3.2.5.3.2. Phase II (Quantitative phase): Structured interviews to test the empirical properties of valuation protocol

After revising the valuation strategy informed by the first phase, the aim of this phase was to quantitatively test the validity of the following psychometric or empirical properties in the application of the proposed valuation protocol.

The recruitment strategy, experimental designs for the selection of SWEMWBS states for valuation, sample size determination, analytical methods to explore the face validity of C-TTO and DCE valuation responses, the feasibility and practicality of the C-TTO and DCE valuation techniques informed by the participants' responses from debriefing questions and the statistics recorded by the EQ-VT software, and the interview process of the structured interview will be documented in detail in Chapter 5.

As this quantitative phase involved the collection of valuation responses from a large sample, econometric techniques were used to model the C-TTO and DCE responses for the generation of preliminary versions of preference-based valuation sets in Chapter 6. A number of criteria for the assessment of model performance across the C-TTO and DCE models will also be described in the model analysis section of that chapter. Moreover, Sensitivity analyses were carried out to inform the robustness of the derived models.

The independent result from the qualitative and quantitative piloting phases will be analysed. A data synthesis matrix representing the main results from both phases will be constructed in Chapter 7 to aid comparison and interpretation.

Chapter 4: Cognitive interviews for the qualitative validation of valuation protocol

*[The results of this chapter were presented and discussed at the 2021 International Health Economics Association (iHEA) World Congress, the Health Economists' Study Group Summer 2021 Meeting organised by the University of Cambridge, and the Warwick Medical School (WMS) Postgraduate Research Symposium at the University of Warwick. Some results have been published in *Quality of Life Research*.]*

4.1. Introduction

Chapter 3 reviewed existing health state valuation techniques and presented the proposed valuation protocol for SWEMWBS. This chapter documents the first phase of testing the validity of the valuation protocol. Specifically, this phase aims to investigate the cognitive process of completing C-TTO and DCE exercises for the valuation of the SWEMWBS. The results of this phase were used to inform the optimisation of the valuation protocol by identifying potential areas of improvements or modifications. The revised valuation protocol was further tested quantitatively in a larger sample in the UK, which will be discussed in the next chapter.

4.2. Methods

Face-to-face cognitive interviews were conducted to investigate the completion processes of the C-TTO and DCE exercises (examples shown in Figure 3 and Figure 4). Participants were asked to think aloud during and after the tasks within a CAPI setting. The tasks were displayed on a laptop screen and participants were guided to select the answers by themselves with the aid of a mouse. This research was approved by the Biomedical and Scientific Research Ethics Committee at the University of Warwick (Reference: BSREC.44/19-20).

Figure 3: The C-TTO task

Figure 3A: Conventional TTO for the valuation of state (2111131) considered better than death

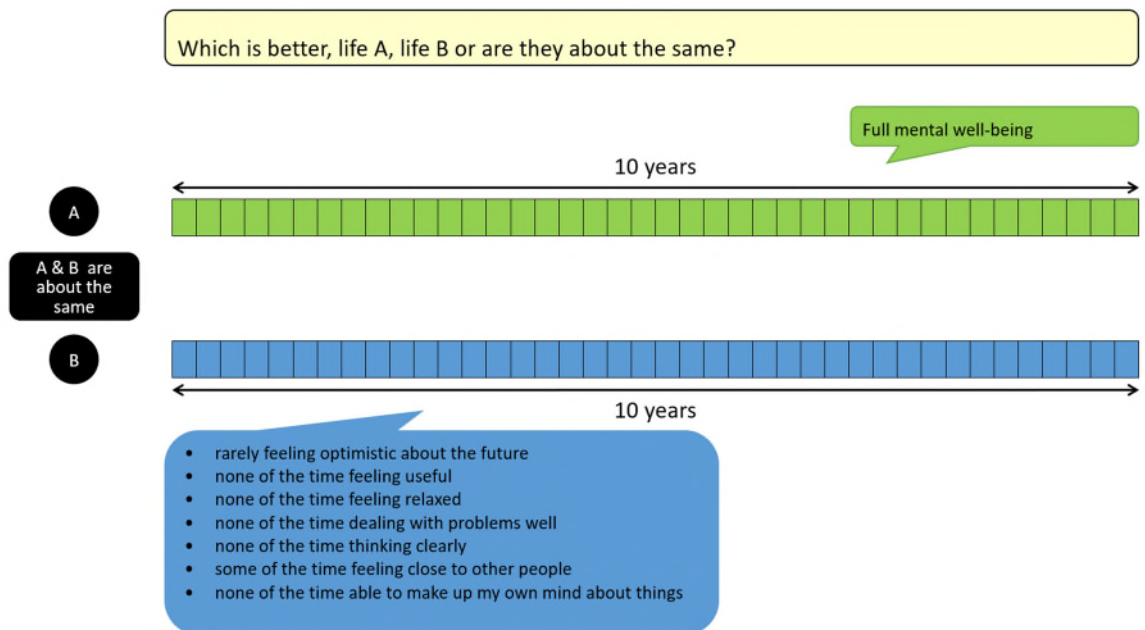


Figure 3B: Lead-time TTO for the valuation of state (2111131) considered worse than death

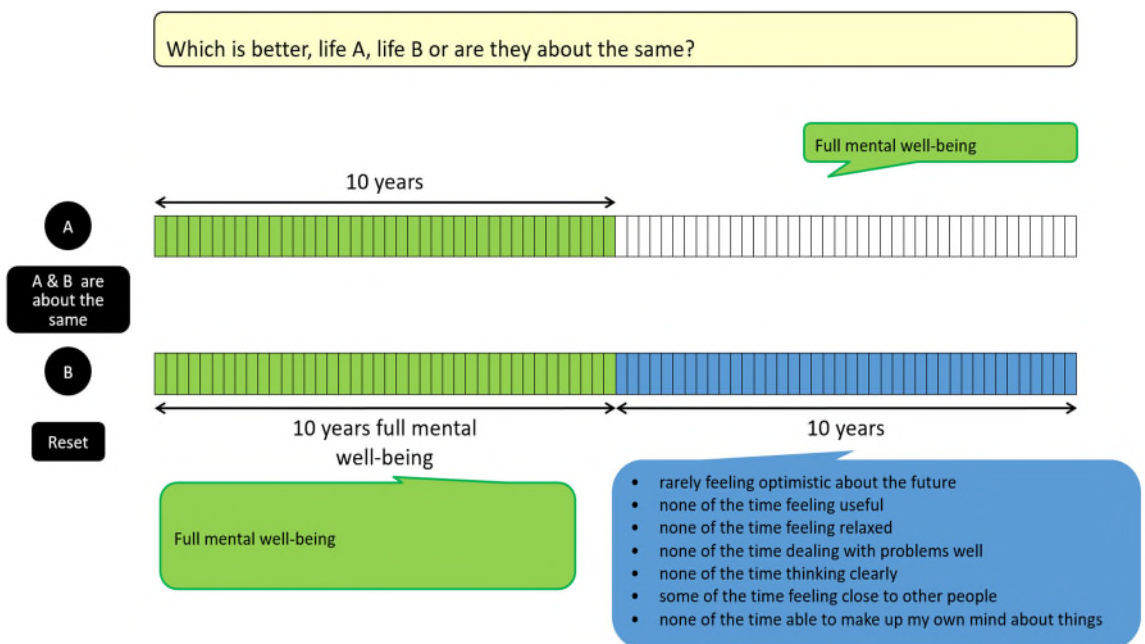
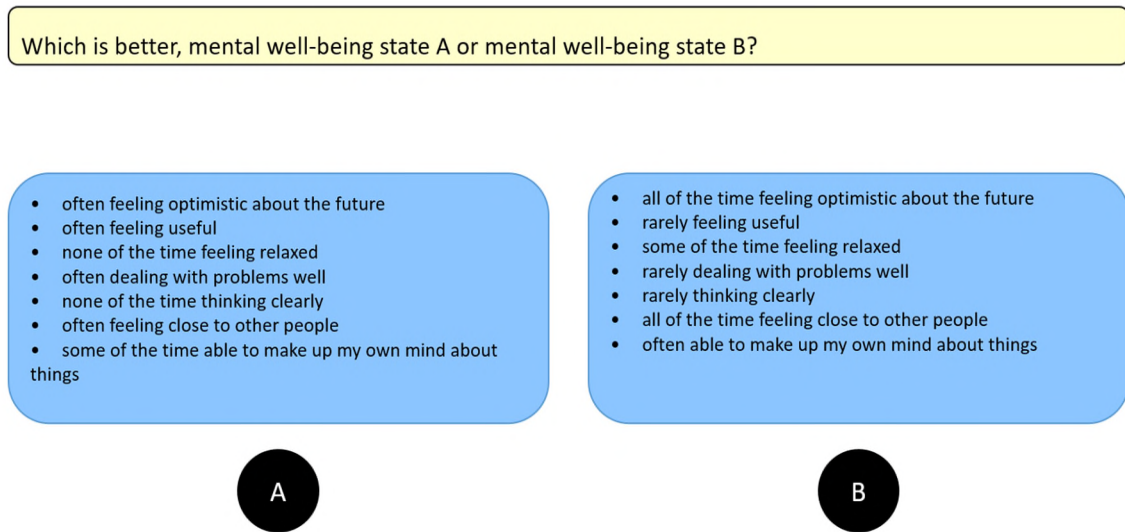


Figure 4: A pairwise DCE with forced choice



4.2.1. Recruitment of respondents

Based on the recommendation from NICE regarding the source of preference data for valuation (The National Institute for Health and Care Excellence, 2013), a representative sample of the UK population is preferred to elicit public preferences. Considering the lack of advertising funding and the onset of the Covid-19 pandemic, a convenience sample of the Warwickshire and West Midlands population aged 18 or above was recruited. Effort was exerted to diversify the demographic background of the participants, in order to gather views from a wider variety of population. As the objective of this phase was to check the clarity of the valuation process, it was not considered necessary to exclude individuals without possession of British citizenship. The main sampling source was university staff and students. They were identified through personal networks. Also, the WMS communication team was contacted to help advertise this project through the weekly WMS Newsletter available to all WMS staff and students. Moreover, this project was announced within the webpage for the Centre for Health Economics at Warwick (CHEW). Basic information for this research project and my email address were provided within the advertisement. An example of the advertising poster is shown in Appendix 11. Interested participants contacted me through emails. To obtain a diverse sample, university students, teachers, administration staff, cleaners, grounds staff were all considered for the sample. The sampling was augmented with invitations to adults from local community groups (e.g. church groups, yoga groups, choirs, or sports teams).

Values were obtained from the sample of general population instead of mental health patients because of several reasons. First of all, the idea of developing a preference-based tariff for the MWALY was to inform resource allocation decision making by publicly-funded healthcare or well-being interventions. Because of this, public utility values elicited by the social decision maker were needed to reflect public preferences. Also, due to the positive framing and generic nature of the SWEMWBS, it was not appropriate to restrict the valuation exercise to mental health patients because SWEMWBS is capable of discriminating a range of mental health states in non-clinical populations. Valuation tasks performed by a broader coverage of the general public would be more meaningful in this sense.

4.2.2. Sample size

Motivated by the principle for specifying data saturation proposed by Francis *et al.* (2010), the initial sample size was set at eight. The interviewer (HHEY) continued to recognise different themes of shared beliefs and the stopping point was applied when there were no new informative ideas identified for three consecutive interviews beyond the eighth interview.

4.2.3. Experimental design for the selection of SWEMWBS states

4.2.3.1. Design for the DCE

As mentioned in section of Chapter 3, a D-efficient design was used for systematically generating DCE choice tasks. As there were no existing preference elicitation articles for the valuation of SWEMWBS and other mental well-being instruments, zero prior parameter values about preferences were assumed in the utility functions. Although efficient design was not superior to orthogonal design in terms of efficiency gain when there was no prior information, efficient design would be more flexible in terms of capturing the effects of different attributes' levels on utility.

The software Ngene was used for the construction of DCE experimental design. Appendix 12 shows the syntax for executing the D-efficient design in Ngene. Given that a pairwise DCE with forced choice had two alternatives within each choice task, the required total number of choice tasks within the design was calculated as below:

$$\text{No. of choice tasks} = \text{round up} \left[\frac{\text{no. of parameters}}{\text{no. of alternatives}-1} \right] \quad \dots (1)$$

No. of parameters = 28 (four dummy variables for each of the seven items) + 1 (the alternative specific constant) = 29

No. of alternatives = 2

$$\text{no. of choice tasks} = \text{round up} \left[\frac{29}{1} \right] = 29$$

Considering a diversity of choice pairs and the minimisation of between-pair variance, it was decided to include more choice tasks within the experimental design before they were allocated into blocks. The efficient design systematically generated 32 choice tasks, which were then randomly allocated into 4 blocks. Each participant was asked to value 1 block, consisting of 8 choice tasks. The result of these 4 blocks is provided in Appendix 13. This design had the lowest D-error (0.41) among all generated designs.

There were different considerations for the sample size calculation of the experimental design. For this qualitative phase, as the aim was to check and understand the valuation protocol, the valuation outcomes were not modelled. Without the need to consider statistical validity, it was sufficient to have a small sample and the decision was informed by the theory of data saturation (Francis *et al.*, 2010). The sample size construction has been described in section 4.2.2.

4.2.3.2. Design for the C-TTO

As mentioned in section 3.2.5.2.2 of Chapter 3, a blocked design was used for the selection of C-TTO states. Each block consisted of two compulsory states [i.e. the lowest mental well-being state (1111111) and one of the closer to full mental well-being states (4555555, 5455555, 5545555, 5554555, 5555455 and 5555554)] plus other states randomly generated by the design. There is no official guidance regarding the required number of states in each block. To comparatively investigate the amount of tasks affordable by the participants, consistent with the DCE experimental design, the number of states in each block was set as 8. In other words, on top of the two compulsory states, each participant was required to value 6 mental well-being states generated using the “AlgDesign” package in R.

The code required to generate this experimental design is shown in Appendix 14. Firstly, a random sample of 42 mental well-being states was generated. In order to check level balance within the subset, a level balance criterion constructed by the EuroQol group for

the C-TTO experimental design was used (Oppe & Van Hout, 2017). The idea was to count the number of appearance of each level-domain combination and to check whether each level of one item appears the same number of times. The value of the level balance check was calculated by the formula below:

$$\begin{aligned} & \textit{Value of the level balance check} \\ & = \sqrt{\textit{sum of squares of the differences between the presence of levels per dimension}} \\ & \dots (2) \end{aligned}$$

The lower the value for the level balance check, the better would be the achievement on level balance, and vice versa. The whole algorithm was running for 10,000 iterations, in order to get a subset with the lowest value in level balance check. Finally, the best subset was randomly and evenly divided into 7 blocks. The additional 8 fixed states (1111111, 4555555, 5455555, 5545555, 5554555, 5555455, 5555545 and 5555554) were set at prior and each block was assigned the lowest mental well-being state (1111111) and one of the randomly allocated closer to full mental well-being states (4555555, 5455555, 5545555, 5554555, 5555455, 5555545 and 5555554). Each participant was asked to value one block.

The results of the 7 blocks generated by R are provided in Appendix 15. This selected factorial design achieved the lowest value for the level balance (22.36), compared to all other generated designs.

A summary of the C-TTO and DCE designs is provided in Table 6 below:

Table 6: A summary of the C-TTO and DCE experimental designs in the qualitative phase

| DCE design | C-TTO design |
|--|---|
| Design: an efficient design | Design: A blocked design with an achieved level balance |
| Total number of choice tasks: 32 | Total number of mental well-being states: 50 |
| No. of blocks: 4 | No. of blocks: 7 |
| No. of choice tasks per participant: one block, consisting of 8 choice tasks | No. of mental well-being states per participant: one block, consisting of 8 mental well-being states [2 compulsory mental well-being states lowest mental well-being state (1111111) and one of the closer to full well-being states |

| | |
|--|--|
| | (4555555, 5455555, 5545555, 5554555, 5555455, 5555545 and 5555554) plus 6 randomly generated mental well-being states] |
| Sample size: informed by the theory of data saturation | |

C-TTO indicates composite time-trade off; DCE, discrete choice experiment.

4.2.4. Valuation platform

The adjusted EQ-VT 2.1 was the most up-to-date protocol with a strict quality control process for recording C-TTO and DCE responses (Stolk *et al.*, 2019). The EuroQol Portable Valuation Technology (EQ-PVT), a replica of the adjusted EQ-VT 2.1, was used throughout the interview and participants completed tasks displayed on the interviewer's laptop.

4.2.5. Interview process

All interviews were audio recorded to ensure the possibility to accurately refer back to the full verbal record whenever necessary. Respondents were interviewed in their homes or at the university campus with the following procedure followed:

- (1) The interviewer introduced the study purpose.
- (2) The participant was asked to sign the consent form indicating the willingness to participate in this study.
- (3) The participant was introduced to the SWEMWBS descriptive system and was asked to complete the SWEMWBS in the Qualtrics survey tool describing their own mental well-being. It was then followed by several demographic questions. A think-aloud warm-up exercise involving 'window counting' was provided to the participant (Collins, 2014).
- (4) The C-TTO exercise: Preference elicitation studies of the EQ-5D-5L incorporate a warm-up example with the imagination of being in a wheelchair as the health state scenario. For the valuation of SWEMWBS, the example needed to be changed to suit the nature of a mental well-being valuation. Instead of a wheelchair scenario, the participant was guided through an example of mental well-being states brought about by being regularly rejected following job applications. The participant was asked to imagine the feeling of lacking confidence and having low self-esteem because of this. The participant was also instructed that full mental well-being is defined as "all of the time" for all the seven SWEMWBS items. With reference to the EQ-VT 2.1, dynamic

questions were added after the first practice example to allow interviewees to become familiar with another evaluation space (Stolk *et al.*, 2019). Similarly, for the valuation of SWEMWBS, dynamic questions regarding the assessment of which state is better (i.e. being accepted for the most ideal job) and worse (i.e. regularly being rejected following job applications, and constantly suffering a poor relationship with friends) than the previous examples were asked for valuation. After these, three practice SWEMWBS states were provided: high (4554545), low (2111131) and intermediate (4212354) mental well-being states. Next, the participant completed the eight valuation tasks, consisting of 2 compulsory mental well-being states [the lowest mental well-being state (1111111) and one of the closer to full mental well-being states (4555555, 5455555, 5545555, 5554555, 5555455 and 5555554)] and 6 randomly allocated mental well-being states generated by the software R for the construction of a blocked design.

To reduce recall bias, during the process of completing the first *three* tasks, each participant was asked to think-aloud everything that came to mind (i.e. thoughts and feelings) concurrently (i.e. concurrent think-aloud).

To save time and reduce the respondent's fatigue, for the remaining tasks, participants were not asked to think-aloud during the process of completion. Each participant was asked to think-aloud retrospectively only after completing all *five* remaining tasks (i.e. retrospective think-aloud). Probing questions in Table 7 were used to complement the interviewee's responses and concurrent and retrospective cognitive processes if they remained inactive.

Finally, the rank ordering inferred by valuations was displayed in the Feedback Module (Figure 5). Each participant was asked to flag any disagreements or inconsistencies with the results, but was not asked to alter the problematic valuations. Some remaining debriefing questions in Table 8 were also asked if they were previously unaddressed.

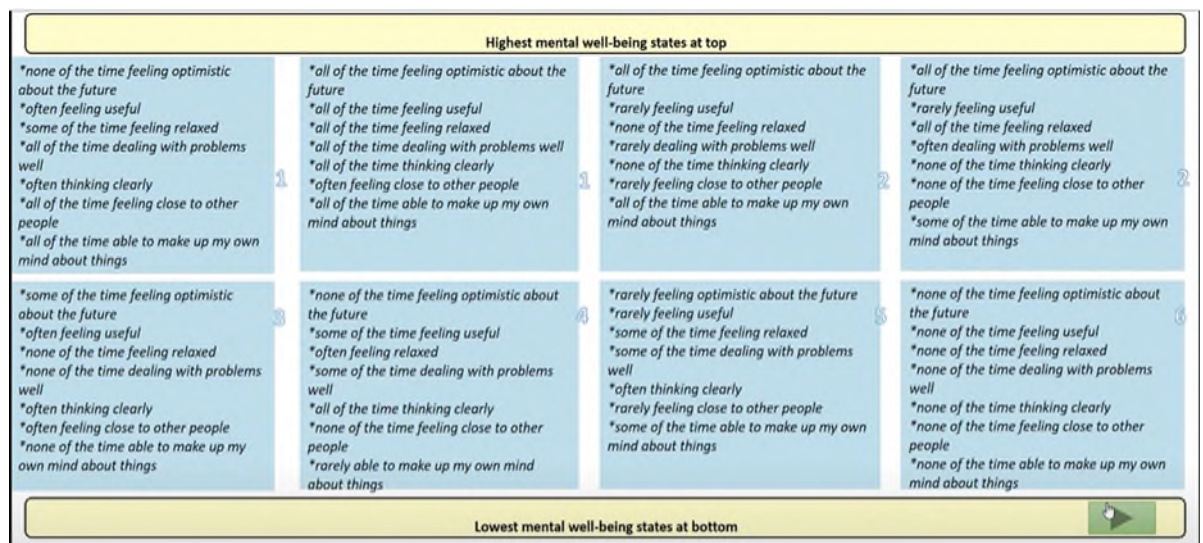
(5) The DCE exercise: Each participant was invited to choose the preferred option among two mental well-being states for each of the 8 pairs of choice task. The paired comparisons and the left-right order of each set of two states were randomised using the EQ-PVT platform. Similar to the C-TTO exercise, concurrent think-aloud and retrospective think-aloud were applied to the completions of the first *three* and remaining *five* tasks,

respectively. These were supplemented by probing questions in Table 9. Some remaining debriefing questions, as stated in Table 10, were also asked.

(6) Finally, the participant was given some overall debriefing questions for both parts of the interview (Table 11) if they were not addressed within the think-aloud process.

(7) The interview was closed with the expression of thankfulness for the participation of this cognitive interview.

Figure 5: A visual presentation of the C-TTO Feedback Module



The Feedback Module displayed the rank ordering of all completed eight C-TTO tasks implied by respondents' valuations. The number appeared at the right-hand side of each blue box corresponded to its rank ordering. The rank ordering of particular states was identical when respondents gave the same trade-off answers to those states.

C-TTO indicates composite time-trade off.

Table 7: Examples of probing questions during the think-aloud process for the C-TTO tasks

| |
|--|
| <p><i>“Could you tell me more about how easy/difficult completing this time trade-off task was?”</i></p> <p><i>“You told me that you felt confused about determining the indifferent point for some of these 8 trade-off tasks, could you tell me more about it?”</i></p> <p><i>“What thoughts came to mind when you were making trade-offs between different mental well-being states?”</i></p> |
|--|

Table 8: Follow-up debriefing questions if they were not addressed within the think-aloud process of the C-TTO tasks

“Were the practice tasks useful for you? How?”

“Did you think the instructions for the practice tasks clear for you?”

“Could you summarise the factors were you considering when deciding the indifferent point?”

“Did you find the number of valuation tasks (i.e. 8 trade-off tasks) manageable for you?”

“Could you tell me how easy/how difficult of completing these 8 valuation tasks were in general?”

“Was the feedback slide useful for you?”

Table 9: Examples of probing questions during the think-aloud process for the DCE tasks

“Could you tell me more about how easy/how difficult of completing this task was?”

“You told me that you felt confused about choosing between this pair of mental well-being profiles, could you tell me more about it?”

“What thoughts came to mind when you were making a choice between this pair of mental well-being profile?”

Table 10: Follow-up debriefing questions if they were not addressed within the think-aloud process of the DCE tasks.

“Could you summarise the factors were you considering when deciding the most preferred option between pairs of mental well-being profile?”

“Did you find the number of valuation tasks (i.e. 8 tasks) manageable for you?”

“Could you tell me how easy/how difficult of completing these 8 valuation tasks were in general?”

Table 11: Overall debriefing questions for both parts of the interview

“Did you think the first part of the interview (i.e. to make trade-off between choices of imaginable life) is easier or more difficult than the second part of the interview (i.e. to look at pairs of mental

well-being profiles and choose the one you prefer)? Or did you feel roughly the same for both parts? Were they still manageable for you?”

“Was the total number of valuation tasks in this interview (i.e. 8 trade-off tasks and 8 choice tasks between pairs of mental well-being profile) manageable to you?”

“Would you prefer to have both parts of the interview or would you prefer only either one of them?”

“Do you have any final overall feedback or comments of this interview?”

4.2.6. Data analysis

After all interviews were completed, verbal information was transcribed verbatim. Thematic analysis was used to analyse data collected by the concurrent and retrospective think-aloud techniques (Braun & Clarke, 2006; Coast, 2017). In addition to summarising verbal text into different themes, the latent level of themes was the main focus in this thematic analysis, to investigate the concepts and underlying ideas of the text beyond semantic level (Boyatzis, 1998; Braun & Clarke, 2006). An in-depth insight of the verbal information could then be obtained to meet the goal of understanding feelings and thoughts of interviewee in cognitive interview.

Firstly, open coding for the first four transcripts was performed by the first rater (HHEY) to identify task completion issues within the text. Coding was discussed and refined with a second rater (HA). With reference to the open coding for the first four transcripts and the field notes for the remaining transcripts, a coding tree for axial coding was then constructed by the first rater. Next, the axial coding framework was applied to code two informative transcripts by the first rater. The second rater coded one of these transcripts and a third rater (JM) coded both transcripts. Upon completion of independent coding for the two transcripts, coding differences were discussed to enhance the consistency and reliability of the coding methods. A more robust version of the coding framework was developed after incorporating feedback raised by the raters. This was applied to code the remaining transcripts by the first rater. Nvivo was used for tagging and labelling potential codes. A codebook to describe the meaning of codes and a descriptive account to re-categorise the coding materials for generating higher-order themes were produced. An explanatory account was finally produced to selectively include quotes for the codes under each higher-order theme (Al-Janabi *et al.*, 2019).

As the aim of this phase was to check and understand how participants feel and think about the valuation protocol qualitatively, the valuation responses obtained in this phase were not modelled. Also, the statistical power of regression models would be very weak given the small sample size.

4.3. Results

Table 12: Demographic characteristics of 14 participants

| Characteristics | Number of participants |
|--|-------------------------------|
| <i>Gender</i> | |
| Male | 5 |
| Female | 9 |
| <i>Age</i> | |
| 18-30 | 3 |
| 31-40 | 5 |
| 41-50 | 2 |
| 51-60 | 2 |
| >60 | 2 |
| <i>Highest education level attained</i> | |
| GCSE | 1 |
| O-Level | 2 |
| A-Level | 2 |
| Undergraduate | 4 |
| Postgraduate | |
| Master | 2 |
| PhD | 3 |
| <i>Ethnicity</i> | |
| White | 12 |
| Asian / Asian British | 1 |
| Arab | 1 |
| <i>Occupation</i> | |

| | |
|-----------------------------------|-------|
| Administrator/Manager/Coordinator | 6 |
| Researcher | 3 |
| Student | 1 |
| Cleaner | 1 |
| Retired | 3 |
| | |
| <i>SWEMWBS score</i> | |
| < 20 | 0 |
| 20-25 | 2 |
| 26-30 | 10 |
| 31-35 | 2 |
| Mean score | 27.64 |

Fourteen interviews were performed to achieve data saturation. The fourteen interviews were conducted between 11th February and 18th March 2020. The interview time was ~60-75 minutes per participant. Table 12 describes the characteristics of participants. Participants highlighted the strengths and limitations of applying the valuation protocol and the completion process. Six broad themes were generated following analyses of the verbal text: format and structure, items and levels, decision strategies, valuation feasibility, valuation outcome, and reflections on mental well-being.

4.3.1. Theme 1: Format and structure

Participants appreciated the well-organised computer setting of the EQ-PVT platform and the automatic allocation of states. However, there were areas for improving the content of the tasks.

4.3.1.1. Inappropriate examples

Despite most participants understanding the C-TTO practice scenarios, two participants pointed to the irrelevance of the job searching example as they were not current job seekers.

“One of the things which I found... difficult or that you might want to change as a point was the focus on rejection of job applications regularly. I don't know I've not applied for a job three years, so it is sort of seem like irrelevant if that makes sense to my day-to-day life... did that make sense?” [Male, 32]

“This is a really tough one... because I'm 67 and I don't really care about job applications.”
[Female, 67]

4.3.1.2. Increase in the variety of preliminary assessments

Participants also suggested the inclusion of an overall health assessment and tasks not related to mental well-being, in addition to the completion of the SWEMWBS to describe their own status. The idea was to explore potential factors of influencing mental well-being and individual choices.

“I'm just surprised that you haven't got... you know... I'm healthy or... I'm not healthy... or I'm relatively healthy, because that would be a factor in there for me as well.” [Male, 67]

Also, one participant suggested the need to allow participants to rank the importance of each of the seven items in SWEMWBS. It could facilitate the understanding of individuals' preference towards a particular state with different combinations of levels of attributes.

“I think it might be useful for you if maybe... you get them to rank just the items... so rather than just saying like... how do you feel about them... maybe rank how importance it is... ... because maybe when you're analysing the data... maybe that will help you like... understand why people chose... why they chose A instead of B for example, based on the items.” [Female, 21]

4.3.1.3. Confusion on scenario completion

The operation of the C-TTO process was sometimes confusing to some participants as it was unfamiliar to most participants. There were three sources of data entry errors identified: mistakenly clicked the life which was not preferred, failed to properly adjust the number of years, and struggled about the displayed meaning of the states. There was also one minor technical issue related to the computer operation. The DCE exercise simply required participants to have a click on the preferred option between two alternatives and there was no data entry problem identified.

4.3.1.3.1. *Mistakenly clicking the non-preferred life*

Some participants were confused about the transformation of their own preferences to appropriate clicks in tasks even if they knew their own preference of a preferred life under each circumstance. It took time for participants to get used to the trade-off procedure and avoid careless moves.

4.3.1.3.2. *Failure to adjust time properly*

Another problem encountered by participants was the uncertainty about adjusting years of full mental well-being to the desired level. They sometimes had an indifference point between life A and life B in mind at first glance, but they struggled to find a way to proceed step-by-step until reaching that point.

“Even the scale is portrayed in a manner that my mind doesn't work. I find it quite strange to... delete and workup to equate a matching valuation.” [Male, 32]

4.3.1.3.3. *Clarification of meaning of a state*

Sometimes it was necessary to monitor the behaviour of participants and explain things displayed on the slides constantly, as participants might get lost about the information on slides. Specifically, it was necessary to remind participants of the ceiling nature of full mental well-being, as they sometimes interpreted the scenario within the state as a bonus on top of full mental well-being.

“even bonus that or no you're happy with this because your self-esteem is high, because you've just got your job. So I assume that I still have... I have full mental well-being there or is that... carrying on there's low self-esteem....” [Female, 33]

4.3.1.3.4. *System operation issue*

One participant interestingly pointed out that he was not comfortable to use Windows software as he had got used to the Mac software. However, this participant understood the tasks and was able to complete the whole interview.

“The technical thing was the computer because... I use an Apple... so everything seems very strange. But apart from that, it was fun once I got the mouse and... that's fun.” [Male, 67]

4.3.1.4. Improvement of presentation layout

4.3.1.4.1. *C-TTO Feedback Module*

Two participants suggested the inclusion of pictures or colours instead of plain text within the C-TTO Feedback Module, to enhance the differentiation of the eight mental well-being states with their corresponding attribute levels.

“Yeah, because I think when I was reading it, trying to read them over the scenarios... then I was like oh... my god I was so confused now, whereas if you have like some sorts of visual... visual representation, it might be easier to follow and you can see... you can

compare them almost, because you will have... so you'll have... um... there's almost like you'd have scenario 1, 2, 3... the 8 different scenarios. And for each eight you have... a list... like a column, and against each of these categories, it will be like a different levels but by colour, so you'll have like a table... summarising the table with colours..." [Female, 35]

"I don't know how you could present it better, I was trying to think with a bar chart would better... with something like would better but I don't know... I don't know cuz a bar chart also... if you had 8 squares with seven bars of different colours say... that would still be a lot of information." [Female, 67]

Additionally, nine participants disagreed with some of their own rank orderings of the eight completed C-TTO tasks. Although participants unanimously acknowledged the importance of reviewing their valuation answers, five participants suggested the possibility for allowing swapping of states after indicating disagreements.

"Yes, it's quite useful. Um... I think maybe you could add something that let you swap the different boxes there... rather than just clicking on ones that you disagree with. Because for mine, I just clicked on two because I wanted to swap them." [Female, 21]

4.3.1.4.2. *Flow of the interview*

Moreover, a few participants found that the order of the interview mattered and suggested the switching between C-TTO and DCE exercise.

"Em... ... you see to me in some ways... that because the bit that we did first is more difficult, I might flip them if it was me. Because then if I've got used to doing the seven things, and I'm choosing different states, I am choosing one against the other... now I have to choose one against the other plus time. It's kind of a build... but... but..." [Female, 67]

4.3.2. Theme 2: Items and levels

4.3.2.1. Contradiction in levels

Eight participants identified non-intuitive combinations of levels of items presented within states. This was a stumbling block to participants' comprehension and imagination.

"I don't understand how you can think clearly but not deal with problems well. If you think clearly, problems should be solved." [Male, 32]

"Often deal with problems well despite the fact that you can't think clearly now, that is strange. And you can rarely make up your mind, now this does not make sense. I mean

how can I only think clearly some of the time and I can't make my mind up about anything, but I can deal with problems well often!" [Female, 67]

"I found those quite interesting... as I say... some of them you know you're feeling optimistic about the future but you're not feeling useful, and to me that was counter-intuitive because if I'm feeling optimistic then... I would be feeling useful." [quali12]

4.3.2.2. Compensation effect

The seven items in the SWEMWBS are to some extent interrelated in the sense that the negativity of items can be compensated by the positivity of other items, improving the overall impression of a state.

"You know you get those support network... that you might not be able to personally deal well with things but people might be able to help you." [Female, 29]

"I guess the one thing which makes me... towards giving B a bit more weight is about the feeling optimistic about the future means that they were not feeling useful or relaxed at the minute, it might be that they'll be doing a new job in six months time and then that might change. This will make you not want to trade off too much time..." [Male, 32]

One participant found that the interpretation of some items could be captured in another item. Specifically, an optimistic outlook about the future could compensate for unfavourable feeling at present.

"Feeling useful is important but if you are feeling optimistic, it's okay because you don't feel useful now. In the future, you will be able to feel optimistic, so you will be able to feel useful, which is part of the optimism." [Male, 32]

4.3.2.3. Overlapping effect

Moreover, participants thought that some items are very similar to each other in context and could be grouped together during the trade-off process under specific circumstance.

"Em... .. I don't see much difference between the dealing with problems well and the thinking clearly. So for me that's... that's one and that's fairly important because I can't be optimistic about the future if I can't think clearly and make decisions. So being able to make up my mind also comes with dealing with problems when I am thinking clearly." [Female, 33]

"And also the ability to... er... deal with problems and think clearly, they come together. For example, to feel relaxed and to think clearly, they go together like... you know fish and chips, you know they seem to go hand-in-hand." [Male, 32]

4.3.2.4. Non-linear effects of levels

Each of the five attribute levels influenced differently to participants' overall impression of a state. As mentioned by two participants, unit changes in attribute levels were not equally valued.

“It's like a sort of a diminishing return... when you go from none of the time to rarely, it is a big jump. But then rarely to some of the time is still quite a big jump. Then some of the time to often is a smaller jump. Then from often to all of the time... it reduces....?”
[Male, 32]

4.3.2.5. Inferiority of top levels

Although full mental well-being is theoretically feasible, one participant rejected the idea of perfection in mental well-being and preferred a dominated alternative without “all of the time” for all seven items (i.e. non-monotonic valuation). The justification was that a maximal well-being state represented a lack of challenging life experience, which was a crucial element of an exciting and balanced life. Also, full mental well-being was considered unrealistic and could imply a lack of awareness or illusionary thinking, the failure to recognise individuals' self-position.

“I really struggled with... the whole concept of full mental well-being, because full mental well-being as described... is too perfect. I don't believe it and I don't like it... I'm a human being, I have ups and downs, that's quite normal and healthy. And it would be really unhealthy to be in this perfect state of mental well-being all of the time because... what's life about?” [Female, 67]

“Often feeling optimistic... seems to mean more healthier than always feeling optimistic. I think always feeling optimistic is... what we call in English Pollyanna syndrome... ... So all of the time feeling useful... interesting. Everybody likes to feel useful, but to be feeling useful all of the time sounds to me like... very hard work. I don't want to be useful all of the time, some of the time I want to be enjoying myself, some of the time I want to be lazy, some other time I want to be doing yoga.” [Female, 67]

4.3.3. Theme 3: Decision strategies

Various decision strategies were found during the C-TTO and DCE valuation processes.

4.3.3.1. Lexicographic ordering

Participants normally put more weight on important items and less for relatively unimportant items when interpreting the overall impression of a state. However, six

participants exhibited a non-compensatory preference, in which they selected a preferred option based on a subset of the most important attribute(s) (Campbell *et al.*, 2006). This violation of the continuity axiom was particularly obvious in the completion of the DCE exercise as they failed to trade-off all attributes when making a final decision.

“They might instinctively [be] going towards option B... just because you're relaxed, you've got people close to you...” [Male, 32]

4.3.3.2. Interpretation of levels

Nine participants considered the existence of extreme levels at the highest end and the lowest end of the response category. They preferred a state with more balanced attribute levels, which were considered preferable for achieving multiple aspects of mental well-being.

“I would go for B because I think A seems more extreme like none none, and then all all, whereas B is... you know only got one all and one none. So it's sort of more middle of the road.” [Female, 29]

Four participants chose a preferred DCE state with a higher level-sum score by counting the number of occurrences of each level in a state. This strategy was also used in the C-TTO tasks to decide the amount of full mental well-being years equivalent to a particular state.

“Em... so just by looking at I see... there's three all of the time, one often, two none of the time, and one rarely... so... it's quite mixed... and it's quite... the rankings are on different extremes as well... whereas for state A... one some of the time... two rarely... one all of the time, em... an often and one none of the time...” [Female, 21]

“So... for this one, the ratings are more next. But I think it's a kind of half half. Because you have two "often". And then one all of the time... and then one some of the time... two rarely and one none. So I think... probably five years but slightly below that, because it's half half positive, half negative.” [Female, 21]

4.3.3.3. Comparison with previous tasks

Some participants might make decisions according to the impression towards the previous state.

“Because of the word friends, it doesn't seem that... the question doesn't seem and the situation doesn't seem as dramatic as the first time. [clicked Life B and Life A changed to

5 years]. So maybe something like this it's... it's... I've got friends but things aren't going so well, but there's more anguish.” [Male, 32]

“It's definitely a better option than the last one. [clicked Life B until Life A changed to 7 years] So it's gonna be more than the seven. Er... say [clicked Life B and Life A changed to 8 years]... is it gonna be more than eight? I'm not thinking clearly and I'm not dealing with problems. So I think I'd stick with a... it's better than the last option, but it's still not... brilliant.” [Female, 37]

4.3.3.4. Personal and external factors

Participants with different demographic background (e.g. ages and occupations), personal judgements and characteristics (e.g. habits, outlook and commitments in life) influenced preferences towards mental well-being states. Table 13 below shows some examples of quotes related to the influence of personal factors on preferences.

Table 13: Quotes related to the influence of personal factors on preferences

| Personal factors | Examples |
|--|---|
| Age | <p>“Sure. I think a lot of my responses are based on... probably on my age because... as I said to you before, the... the feeling as... we age... em... the feeling of not being able to do things would really... I've quite bother me. Em... I go to the gym twice a week still... walk a lot, with trying keep healthy. And... not to do that... will worry me.” [Male, 67]</p> |
| Life habit, life outlook and commitments | <p>“Thinking clearly you see that's a nice thing to have... but... I'm not sure that's top of my list... I like to daydream... come on anyways so...” [quali02]</p> <p>“Feeling close to people, I don't know... this one is a bit different for me because... I've lived in a lot of countries, and I've moved a lot. So... this one is a bit tough for me because I tend to make friends and for whatever reason they leave I leave. And so it's quite common for me to... not have to... or... I shouldn't get attached to people because it will just mean it will cause a problem later on. So this one, I don't put too... too much importance on it but it is important. No friends at all would be a disaster.” [Male, 32]</p> <p>“Again I touched on the fact that feeling useful em... both in personal life and... would have a big effect on me, when handled my work life.” [Male, 28]</p> |

| | |
|-------------------|--|
| Personal judgment | <p><i>“Um... I don't think so because I think different people prioritise the different categories differently. So maybe if... em... I don't know... if that was all of the time for this one and then none of the time for that one, they would choose this one because... they think that's more important... or something like that.” [Female, 21]</i></p> <p><i>“Em... ... yeah it's really hard because when there's something that is so... you know how do you quantify how much your life you would give up for that. Em... so it's all just personal opinion isn't it.” [Female, 29]</i></p> |
| Personal trait | <p><i>“Em... feeling close to people, em... I think... I'm not really what you would call... the understand the people person. So... I'm quite happy to be by myself so... is possibly less relevant to me than some other people.” [Male, 67]</i></p> |
| Personal belief | <p>A participant believed that things might improve even if there was no indication of any change in item ingredients within a specific period in the state.</p> <p><i>“I work on the... I work on the principle that things might change. You might be feeling about as rubbish as you can, but things do change. When you're not feeling optimistic, said the feeling optimistic was not... I think things come out completely out the blue, and you can feel optimistic about things and then the rubbish gets pulled when you... ... or you can be feeling really rubbish about everything, you're not getting a job and all of a sudden you'll get three interviews at once... so... I think optimism is sort of how you choose to look at what's going on around you rather than... that the world is against you.” [Female, 37]</i></p> |

Furthermore, the existence of external support would increase the acceptability of a particular state. The impression of a seemingly worse state could be improved when there was sufficient help available.

“When I'm thinking personally, do I have... issues dealing with problems for example. But then I've got a support network around me... ... then I think I'll actually know what... when I've had issues dealing with problems at the fact that I've had a support network around me... is really help. So that's why I think those two things... probably optimistic about the future got to my own current circumstances...and support network around me because that is why I have now. And I thought that is something that's very important...

yeah. I could... it was thinking back to my own family and my own friends and was thinking... oh you know I've had times where I've not really felt very useful but then someone said to me, oh no don't be silly.” [Female, 35]

*“Possibly I don't make up my mind about things, I'll leave things to her (i.e. his wife)...”
[Male, 28]*

4.3.3.5. Availability heuristic

Eight participants assessed the frequency of a class or the probability of an event by the ease with which instances could be brought to mind (Blumenthal-Barby & Krieger, 2015; Kahneman *et al.*, 1982; Tversky & Kahneman, 1973). In other words, they sometimes relied on existing things that came to the mind instantly to make immediate judgements of an event. They explained their impression of a state by recalling daily examples (e.g. news reports and relatives' experiences) and past experiences.

“I think if I'm relaxed, and I feel close to other people, and I often think clearly, and I deal with problems well, then there must be some optimism about the future. Em... but I'd rather not be optimistic about... okay, Brexit is a good example. I'm not optimistic about the future, but I'm quite relaxed.” [Female, 33]

*“I have a brother-in-law... .. who had a stroke when he was... late forties... .. so I think this might kind of almost describe him. Because emotively he's still there, but physically... he's not... able to do anything and mentally, he's not able to be doing... very much.”
[Female, 67]*

“I just feel I have been so low in the past, I don't want to ever go back to that. Now I don't want to ever go back to feeling that low, and I don't think it was particularly good... em... from husband this time, and when I had Chris, our youngest, I had postnatal depression. And obviously it has an impact... on the children. Chris is one of our two stepchildren and you know... it was a lot for everybody stripping through life after years, and a blended family and step family at that time.” [Female, 51]

Four participants used an analogy to illustrate the meaning of a state.

*“Not able to make up your own mind at all..... again that's a bit like... being in a prison or institutionalised or something if you can't ever make any decisions for yourself...”
[Male, 32]*

“There's almost as if you're asking if... if you're in this situation, would you rather be in this situation or will be like... almost like suicide.” [Female, 35]

4.3.3.6. Duration of C-TTO states

The impression and valuation of a mental well-being state would sometimes depend on how long to live for a particular life.

“I don't know because... are you supposed to be feeling that for the whole of those 10 years? You often feel optimistic about the future...is that... is that for the whole of those 10 years? So it can't change?” [Female, 51]

Even though a few of them discussed the optimality of the duration of a state, no participant disfavoured with the theoretical setting of 10 years of the allocated state within the C-TTO tasks.

4.3.3.7. Satisficing heuristic

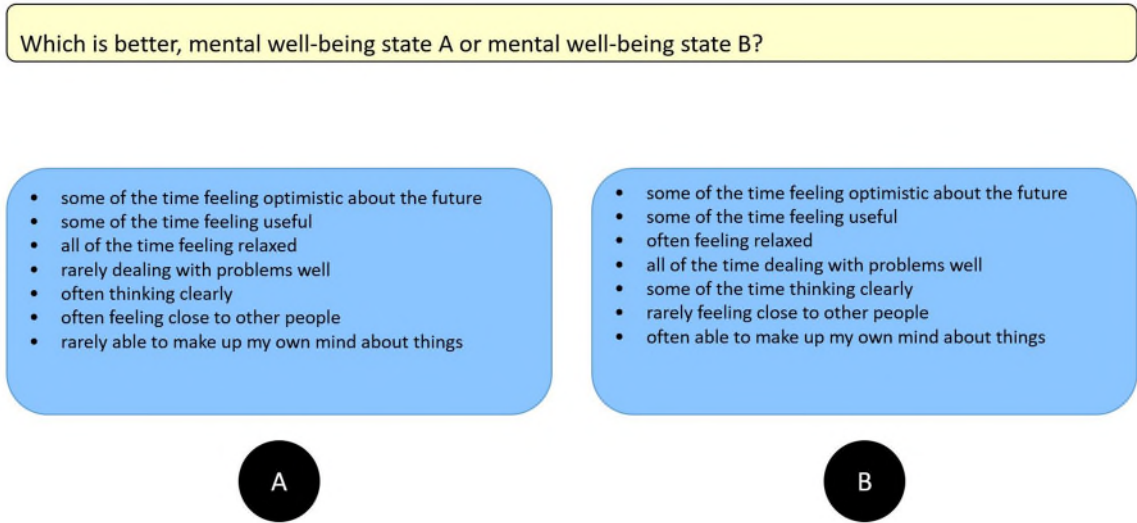
Satisficing heuristics means that an individual ceases a decision-making process when an adequate rather than an optimal solution is reached (O'Sullivan & Schofield, 2018). Participants might make a decision which was not necessarily the best, but at least it was sufficiently fine.

“I'm gonna go for B because... they've got some of everything, except for feeling relaxed... ... and whereas on the other side [state A], they rarely feel relaxed anyway... suppose there's not really much of difference in rarely feeling relaxed [in state A] and none of the time feeling relaxed [in state B]... ... okay I will go for that one [state B].” [Female, 35]

4.3.3.8. Ignorance of identical levels of attributes between DCE alternatives

Some participants agreed that it could be cognitively less challenging for the DCE tasks if the levels of some items across both alternatives were fixed as the same. An example of this is shown in Figure 6 below.

Figure 6: An example of a DCE pair with identical levels for two items



“Yeah, I think that would make it easier. Definitely be easier. It is two less things to worry about. So you are sort of focusing on the remaining bit, does it make sense?” [Male, 32]

“Participant: I don't think it will make a difference. Because if these two are the same, then you wouldn't need to consider them. So I guess you wouldn't need to put it in the scenario anyway... because they would be...”

Interviewer: You will just ignore these?

Participant: Yeah, I would just ignore it because I know that they're gonna be the same. So I would only consider the ones which are different.” [Female, 21]

4.3.3.9. Rejection of unimaginable states

One participant observed that their decision to select a particular state within a DCE pair was sometimes informed by the elimination of an unimaginable state.

“Sometimes I was choosing the other one, not necessarily because I preferred it, but because I rejected one. It's like I just don't believe that.” [Female, 67]

“I rarely feel optimistic about the future... ... at least this is conceivable, I can conceive of A, therefore I will choose it. I don't know that... I don't know that I even prefer or don't prefer it but I can conceive of it [laugh].” [Female, 67]

4.3.3.10. Framing effect

Framing effect is a type of cognitive bias in which the decision to a particular choice varies according to different presentations of information (O'Sullivan & Schofield, 2018). Some

participants adopted different decision strategies for the completions of C-TTO and DCE tasks. The weighting of items were affected by how the tasks were presented and the combination of other levels of items. Also, the trade-off decision might change when it comes to reality rather than imagination.

“That's all sort of feeling like... doing it... here might be quite different to do in practice well it's actually like you will be killed in three years if you do you know what I mean... and if the death sentence was real... ... then you were being killed in four years if you... choose the good state...I might change my decisions slightly but it's... yeah it was tough but very interesting.” [Male, 32]

4.3.3.11. Integration of self-written notes

A few participants raised the idea of drafting some hand-written notes for summarising the information of a state and assisting cognitive analysis.

I probably draw a little table... with... I mean it probably would be very similar to this but... just writing it down helps me to think clearer, so that's sort of again, that's a very personal thing. But yeah... I'm reading off the screen... isn't my preferred I prefer to do things on papers. Very old school. [laugh]” [Female, 43]

“It wasn't such a current dry picture as such... it was sort of having to really you know... I guess I could made the... if I had time and I had a piece of paper, I could really go number on... and how then that's it... and then I would be able to make like... don't know what sort of judgement you call it but... the mathematical judgement... here is my four priorities and then the numbers might say right actually you shouldn't go for that one... but I might made a judgement just again... my snap judgement without thinking-aloud where I should go with that one so again if I really broke them down into my thinking about numbers, then I might... read the other ways...” [Male, 28]

4.3.4. Theme 4: Valuation feasibility

Difficulties such as imagination of states and quantification of years in the C-TTO tasks accentuated cognitive burden. Some participants also felt overwhelmed when completing forced DCE pairs as the process of comparing alternative permutations of levels for seven attributes induced information fatigue.

“It was tough... but... doable... in terms of... used quite a brainpower... it's just you're trying to hold a lot of things in your mind at the same time as you've got the profile of

attributes on the left and then the profile on the right, and then is just trying to weight those up simultaneously.” [Male, 32]

However, encouragingly, all participants found the interviews manageable and the C-TTO and DCE tasks complementary. Participants also acknowledged the importance of the C-TTO practice tasks to relieve uncertainties from mere description of instructions.

“Yes. Because it almost kind of gets you... because I think if you start jumping into the actual exercises, em... they might not be so clear at first. But when you do a couple of practice exercises, they help you like... once you do the real, you start to do the thing you think... okay this is. I know how to do this now. Definitely good kind of practice exercises.” [Female, 35]

Also, the C-TTO practice tasks could help participants recognise their standard and position on time preference. Answers would be more precise in terms of assigning the appropriate number of trade-off years for a particular state. For example, one participant was giving a higher value for the third practice state than the first one, which was regarded as illogical as the third state (issues with job applications and relationships) was theoretically much worse than the first one (issue with job applications only). After discussing his answers, he understood that the third state was poorer with more negative issues going on. He realised that he did not adjust the value properly in the first one. This finding suggested the function of these tasks as a way to allow participants understanding their trade-off preferences and increase the manageability towards the trade-off process.

“Interviewer: But for the first one, you just being rejected for job applications, but maybe you don't have any problems with your friend.

Participant: Maybe for the first one I was a little too... overzealous... .. maybe I was... too binary for the first one.” [Male, 32]

In addition, the importance of incentives was identified as the key element influencing valuation feasibility. Sufficient incentives were undoubtedly vital for participants to ensure the level of engagement in tasks. It was notable that the level of engagement could be enhanced through two channels: Extrinsic and intrinsic motivations. Extrinsic motivation was driven by various forms of external rewards (Deci, 1972). Among all interviews, when investigating the optimal total number of valuation tasks allocated to each participant, one participant raised an issue regarding the difference between willingness and ability to complete. The actual number of tasks capable for completions

was determined by the willingness to complete. The availability of money and gifts was indeed one of the factors boosting the incentives to contribute more to this interview.

“Participant: I think I would... be able to happy to do... maybe like in total 25... maybe 12 for each part... you know... or maybe even 30... it wouldn't be fun though so... ...I think there is a balance of what am I physically capable of doing or... versus what am I happy to do... does it make sense?”

Interview: There is a difference between willingness to do... and what you actually can do.

Participant: I guess some people... well... if I have been paid to do that... I'd be willing to more maybe... .. em... how many... in terms of... yeah so... 25, 30 in total... If you are going to do think-aloud and... yeah... just give people coffee and biscuits...” [Male, 32]

Furthermore, the level of engagement was also affected by the existence of intrinsic motivation, in which a person is intrinsically motivated if he performs an activity for no apparent reward except the activity itself (Berlyne, 1966; Deci, 1972; Hunt, 1965; White, 1959). Interesting and attractive elements should be included within the interview to sustain the participant's attention.

“Interviewer: Could you do more?”

Participant: Em... I could do more... depend on what you were asking me. If you were asking similar things, it becomes... I suspect my level of engagement... would drop. [laugh]” [Male, 67]

4.3.5. Theme 5: Valuation outcome

4.3.5.1. Failure to reach the C-TTO indifference point

One participant with prior experience of mental illness failed to reach the indifference point for four states even after exhausting all lead time in the worse-than-death scenario. Particularly, for the lowest state 1111111, she found it distressing and was not willing to live in this state, no matter how many years of lead time were given ahead of this state. This constituted the value of $-\infty$.

“I think... in this reality where none of those... I don't... I don't know. I feel very negative but I would never want to put anybody through... I never want to go through ten years of feeling like that... so I don't think there's any maximum number of years...” [Female, 51]

Among those participants who valued states as better-than-death, one participant failed to reach the indifference point for some states due to her dislike of the concept of full mental well-being. The task failed to proceed as it violated the theoretical assumption of setting full mental well-being as the best state. This implied a value of >1 .

“Interviewer: The task will not move on if you click B indeed... .. because there's no extra life...”

Participant: Oh, I am not allowed to reject full mental well-being, hey! [laugh]” [Female, 67]

4.3.5.2. Non-trading effects

Nine participants were not willing to give up years of life if the states were considered sufficiently promising. They valued some imperfections in life.

“I think that's pretty good... em... and I feel that life would be boring if it didn't come with challenges and you can't appreciate the ups when you don't have the downs, so for me they're both the same.” [Female, 33]

“I would be happy with either of those, because none of them are particularly... gonna make you sad, are they? Okay, it's not... all of the time, but I don't think life in general is like that... ..” [Female, 51]

4.3.6. Theme 6: Overall Reflections on mental well-being

Participants thought that the C-TTO and DCE tasks within this interview were beneficial and allowed them to reflect more on life and personal preference. It was encouraging to discover that participants recognised the valuable elements of this interview as the process allowed them to realise their cognitive interpretations and attitudes towards different mental well-being scenarios.

“It's interesting. It's good. I think it makes you... reflect more, and like... it makes you actually think about... what's important to you and reflect on how you feel about this now as well.” [Female, 21]

“It's interesting to retain the time to think about these things and think about my mental well-being em... what things I really hold... dear I've never really took the time to em... really think about it before so... it's useful to know that em... but yeah it was enjoyable, I hope you... I hope I've proved useful parts of myself...” [Male, 28]

This interview also helped participants understand more about a theoretical state, which was ambiguous in the past.

“But specifically to the remaining five, I felt that the remaining five were... like I think for the first three, I chose ... I chose... I feel the same questions were being asked the whole way through and I felt consistent. And I felt that maybe understood a little bit more as to what I want... a perfect... perfect mental state to be.” [Male, 32]

4.4. Discussion

This Chapter summarises the issues identified through the cognitive process of completing C-TTO and DCE tasks for the valuation of the SWEMWBS. Implications for modifications (Table 14) and other interview findings are discussed in this section.

Table 14: Issues identified by the interview and the corresponding proposed modification to the valuation protocol

| Issue identified | Related section | Proposed modification of the valuation protocol to be used in the quantitative phase |
|--|------------------------|---|
| Inappropriate C-TTO practice examples | 4.3.1.1 | One additional version of practice example related to physical health and relationship. |
| Confusion about the time trade-off procedure | 4.3.1.3 | <ul style="list-style-type: none"> ▪ More detailed explanations of the instructions. ▪ Slowing the instructing speed. ▪ Encouraging participants to raise questions. ▪ Clarification of practice states before completion. ▪ More step-by-step trade-off demonstrations. |
| Visual difficulty in differentiating the states within the C-TTO Feedback Module | 4.3.1.4 | Guidance to enhance the readability of the states line-by-line was provided. |
| Incomprehensible combinations of levels of attribute | 4.3.2.1 | The selection of experimental design choice sets with potential uncommonly reported states could be avoided. |
| The exhibition of lexicographic ordering | 4.3.3.1 | Participants were instructed to consider all attributes within the allocated states. |

| | | |
|--|----------|---|
| The existence of preference heterogeneity | 4.3.3.4 | Advanced modelling techniques with the inclusion of covariates and interaction terms could be applied. |
| Visualisation of states from a third party perspective | 4.3.3.5 | Participants were told by the instruction to imagine themselves being in the allocated states. |
| Uncomfortable trade-off process for pen-and-paper participants | 4.3.3.11 | Participants were told to use pieces of paper and stationery optionally for integrating self-written ideas. |
| Motivation of participation | 4.3.4 | <ul style="list-style-type: none"> ▪ Improvement of interview design. ▪ Expression of participation thankfulness by money reward. |
| Promising manageability of the number of tasks | 4.3.4 | The number of tasks for each of the C-TTO and DCE parts was increased from 8 to 10 (i.e. 10 C-TTO and 10 DCE tasks). |

C-TTO indicates composite time-trade off; DCE, discrete choice experiment.

4.4.1. Format and structure

Firstly, the style and structure of the interview were challenged. Some participants suggested the increase in the variety of assessments in addition to the completion of SWEMWBS to describe mental health status. It was true that the assessment of an overall health can capture a wider picture of the health status of an individual. The inclusion of some non-mental well-being related C-TTO practice tasks could also facilitate the understanding of the trade-off meaning between choices. However, these were irrelevant or offered little insights to the interview, as the nature of the interview was to gather valuation issues specifically related to mental well-being. In this sense, the completion of SWEMWBS before the valuation exercise was considered sufficient in realising the mental well-being status of an individual. The inclusion of a window-counting think-aloud warm-up exercise, and the six practice C-TTO examples were also adequately covering the concepts required by the participants for the interview. There was no need to include more preliminary tasks. Regarding the possibility to allow the ranking of importance of the SWEMWBS items prior to the completion of valuation tasks, it was not considered in this phase. The reason was to avoid the risk of introducing bias to the valuation process by priming respondents to think lexicographically.

Apart from that, it was worth investigating whether it was essential to tailor different practice examples to participants with different background. Considering the cognitive burden from imagination, one additional version of generic practice example related to physical health and relationships (Appendix 16) was added to the original versions of the job application and relationship examples in the follow-on quantitative phase. The relationship portion of the examples was not altered as no problem of imagination was identified during the interviews in this phase. Participants in the quantitative phase were given the flexibility to choose between two practice versions. Also, inexperienced participants unintentionally made mistakes even after practicing because of the complexity of the C-TTO completion. The presentation context was improved in the quantitative phase by deepening and slowing the instruction explanations. Clarification of the meaning of the life A and life B scenarios after each move was described, ensuring that participants recognised the trade-off purpose.

Moreover, regarding the difficulties in interpreting the C-TTO Feedback Module, more guidance on reading the pooled states line-by-line was provided in the quantitative phase to enhance the visual readability of the C-TTO Feedback Module. This slide was useful to check the robustness of the results as more than half of the participants flagged problematic rank ordering of states. The modelling results from other country-specific EQ-5D-5L value sets with the adoption of EQ-VT protocol also showed a goodness-of-fit improvement after dropping flagged states (Ferreira *et al.*, 2019; Wong *et al.*, 2018). Some participants suggested corrections to rank ordering deliberately by allowing swapping between states. However, arguably, this would sacrifice the role of C-TTO in deriving the value of states. To keep the C-TTO theoretical foundation, with reference to the EQ-VT, data from those flagged invalid states should be deleted and no swapping of states was required after indicating disagreements (Stolk *et al.*, 2019).

Furthermore, it was worth noting that the order of the interview flow was subject to individual's difference in taste and preference. There was a suggestion of putting the C-TTO part after the DCE part. However, the structure of the C-TTO tasks was more complex in terms of instruction information and completion procedure. If the C-TTO tasks were put in the second part of interview, participants might arguably lose patience in digesting the information after going through a number of tasks in the first part of the interview. The concentration or attention level of the participant could also deteriorate

when participants moved towards the second part of interview. Thus, the motivation and sense of engagement in the C-TTO tasks could be reduced. Considering all participants agreed that all tasks presented under this current flow were manageable, there was no indication of an essential need to change the order of the interview in the quantitative phase.

Lastly, for the potential difficulty of using the Window operation system faced by regular Mac users, this issue was considered very minor. I was confident that participants would be able to complete the exercise as long as a clear instruction regarding the appropriate buttons for each move to click was provided.

4.4.2. Items and levels

Some features of the valuation items and levels were identified by the interviews. It was normal to discover different forms of interaction effects between the SWEMWBS items, as items were supposed to be correlated with each other for a scale to be considered coherent. The confirmatory factor analysis of the SWEMWBS was conducted across different settings in the U.K. (e.g. Haver *et al.* (2015); Vaingankar *et al.* (2017)) and the uni-dimensionality of measuring mental well-being was suggested. This empirical property supported the qualitative finding in this phase, in the sense that items were supposed to have certain degree of correlations if they were measuring the same underlying construct. As a matter of fact, it would be worth investigating whether the incorporation of all second-order interaction terms between levels of attributes (i.e. $ITEM1_{L1} * ITEM2_{L1} + ITEM1_{L1} * ITEM2_{L2} + ITEM1_{L1} * ITEM2_{L3} + \dots$) in addition to the main effect parameters would improve the fit of the model. However, the complexity of the model would be largely increased as SWEMWBS has 7 items with 5 response categories each, which could generate additional 140 parameters within the modelling specification. It would also be difficult for experimental software to find an efficient design for the DCE exercise. Also, a relatively large sample size (e.g. at least ~350) would be required to sufficiently model the utility function. It might not be feasible under tight time and resource constraint within this PhD. In this context, the testing of second-order interaction terms was not the focus of the quantitative phase. Nevertheless, it should be noted that future SWEMWBS valuation study with a sufficiently large sample size could analyse the issue of second-order interactions on the reliability of estimated parameters.

Regarding potential conflicting combination of levels of attributes within a state, it was worth investigating the need to exclude potential implausible combinations of levels of attributes out of the 78,125 SWEMWBS states. This could improve the valuation experience by reducing the chance to encounter states that some participants might consider unconscionable during the completion of the valuation exercise. In this context, national datasets in the U.K. that include the SWEMWBS were separately analysed to explore characteristics of response patterns to the measure (Appendix 17). The purpose was to identify highly uncommon combinations of levels of attributes, so that they could be avoided when allocating the choice set given to the participants. Interestingly, there was insufficient evidence to exclude any SWEMWBS states as the implausible states claimed by participants were not uncommon in national survey responses. Instead of state exclusion, when allocating choice sets to participants in the quantitative phase, the selection of experimental designs with potential uncommonly reported states could be avoided among many iterations.

Moreover, regarding the non-linear effects of attribute levels, dummy coding (i.e. inclusion of 28 dummy variables) was used in the utility specification of the DCE experimental design (Daly *et al.*, 2016). The interview results supported this assumption, as some participants indicated that they placed different weights on different levels. When the level of an attribute was above or below a specific response category, that attribute would become more or less important. This implied that the effect of different levels on the utility was not always the same.

Lastly, no valid conclusion about the issue of non-monotonic valuation (i.e. not preferring full mental well-being) on the suitability of C-TTO technique can be made as this was only identified by one participant. The C-TTO valuation technique was still used in the quantitative phase, with a view to investigate the proportion of participants exhibiting this kind of non-monotonic valuation.

4.4.3. Decision strategies

Additionally, completion heuristics were discovered. The presence of lexicographic ordering, a focusing effect normally discovered when respondents interpreted a state (Ryan *et al.*, 2009), caused the failure to reflect full preference when some attributes were unattended. Moreover, the strategy of solely interpreting the level-sum score of states within the DCE pairs posed a risk of neglecting the essence of items. In other words,

participants might focus only on the levels within a state, without keeping their corresponding items in mind. Considering these, participants were reminded to interpret a state with both its levels and attributes before task completions in the quantitative phase.

It was discovered that the answers to some valuation responses were induced by referencing the answers to the previous tasks. This implied that the value given for each choice task might not always be independent. This phenomenon was more common in the C-TTO responses, in which the quantification of a state was sometimes influenced by the values given to the previous tasks. Undoubtedly, the completion of C-TTO practice tasks was important for participants to figure out their own quantification standard through learning effects. The reliance on the past values could then be minimised.

Moreover, it was noted that the duration of a C-TTO state could influence the trade-off decision of participants. As there was no indication of inappropriate specification of the time horizon of both life A and life B within the C-TTO tasks, there was no change to the duration of mental well-being state when rolling out to the quantitative phase.

Furthermore, the demographic background of an individual influenced the interpretation of a state and altered the weighting of items. Particularly, values attached to a specific state were influenced by the variation in individuals' characteristics and tastes. This was a form of preference heterogeneity, in which the coefficients of the attributes across individuals might not be constant. Choice models can explain deterministic (across observed individual characteristics) and random (unobserved) heterogeneities (Lancsar *et al.*, 2017). As the C-TTO and DCE responses would be modelled to generate utility values in the quantitative phase, advanced modelling techniques were considered in Chapter 6 in addition to the main effects models. Also, the existence of preference heterogeneity suggested the importance of including a sample with diverse and representative demographic characteristics, ensuring the capture of well-rounded opinions by averaging the views from the general public.

Some participants raised the possibility that their trade-off decisions to these inexperienced scenarios could be different when it comes to the reality, i.e. hypothetical bias. This was indeed a limitation that a hypothetical trade-off decision could not completely substitute the decision making in reality. However, this was also a limitation of other existing health valuation techniques. It was difficult to ensure that participants

were allocated with states that they had experiences on. Also, a few participants visualised states through the lens of available examples in society or through a third-party state. Participants were reminded in the instruction of the quantitative phase that the theoretical setting of both C-TTO and DCE techniques required them to primarily immerse *themselves* into the allocated scenarios, rather than imagining how others would behave in the state.

It was realised that keeping several levels of items identical between the two DCE alternatives could have relieved participants' cognitive burden. However, as documented in other studies which tested the effect of overlapping some dimensions across pairs (Mukuria *et al.*, 2021), participant's neglect of these identical items made the trade-off decision less informative. It was interesting to note by one participant that the selection of preferred option within a forced DCE pair was sometimes informed by the rejection of unimaginable state. Together with the exhibition of satisficing heuristic as a buffer strategy in making trade-off decisions, these strategies implied that the chosen state might not always represent a preferred state or a state which was comfortable to live with. It could be just a relatively acceptable state. In any case, to help reduce participants' chance of encountering unimaginable states, the proposed idea of avoiding highly uncommon states (Appendix 17) should be adopted.

Finally, it was understandable that not all participants could get used to or think effectively under a CAPI setting. Considering that participants might need pens and papers to assist their cognitive organisation, participants were told to use pieces of paper and stationery freely during the quantitative phase.

4.4.4. Valuation feasibility

Concerning the manageability of the exercise, the function of the C-TTO practice tasks in boosting manageability of the trade-off process was obvious. Initial learning of the trade-off quantification was required to ensure precise adjustment of the years of full mental well-being in life A. The inclusion of a practice section for the DCE exercise in the quantitative phase was deemed unnecessary because the decision button was simple to operate. Participants were only required to apply a one-step process of clicking a button of the preferred option. A clear introductory instruction was believed to be sufficient for participants to follow with no difficulty.

Moreover, even though some participants felt cognitively exhausted to answer a forced DCE pair, the idea of a forced choice was to maximise the trade-offs between items and avoid the loss of power (Veldwijk *et al.*, 2014). The possibility of including an opt-out alternative was not considered in the quantitative phase. The idea was to avoid losing significant amount of preference information when participants constantly relied on the opt-out option to bypass the trade-off difficulty. Actually, it was not bad to hear that participants struggled about trading off between similar DCE alternatives. This indeed revealed the strength of the experimental design. The software Ngene systematically generated a wise set of choice tasks that allowed participants to experience trading off across various combinations of levels of attributes. The tasks would be meaningless if the differentiation between states was too obvious. Also, it could sometimes be difficult for participants to compare alternative permutations of levels for seven attributes. However, it was considered impossible to further reduce the number of items as the SWEMWBS descriptive system has already undergone comprehensive Rasch analyses (Bartram *et al.*, 2013; Stewart-Brown *et al.*, 2009). The 7 items incorporated in the SWEMWBS were considered to satisfy the uni-dimensionality of covering a single latent theme of mental well-being. As mentioned in section 4.4.3, it was also unwise to fix some levels of attributes across the alternatives, due to the risk of negligence during the trade-off process.

In addition, interestingly, the imagination concern regarding the unrealistic full mental well-being state was not supported by the existing national data available in the U.K. As stated in Appendix 17.1, among the pooled Understanding Society and Health Survey for England datasets, the state 555555 (full mental well-being) was regarded as the third commonly reported SWEMWBS states out of the 78,125 states in total. In the context of its high frequency of reporting, it was argued that full mental well-being was indeed a practical state in reality. It was not really exhausting for the general public to imagine this state in life A of the C-TTO exercise.

Furthermore, appropriate incentives were required to motivate participations. The sample recruitment for the quantitative phase was challenging as a large sample size was required. To increase the speed of recruiting sufficient number of participants within the tight time schedule, promising strategies should be used to increase public motivation to participate in the valuation interview. Strategies to boost intrinsic motivation were discussed in section 4.4.1. These included the incorporation of more user-friendly characteristics of the

interview through improving presentation layout, increasing the choices of practice examples to address imagination burden, etc. In terms of extrinsic motivation, the level of engagement could hopefully be raised when participants were given money shopping vouchers.

As all participants found the number of tasks within the interview manageable and a majority expressed the ability to complete more tasks, the number of tasks for both the C-TTO and DCE in the quantitative phase was increased by two each, i.e. each participant was asked to complete 10 C-TTO tasks and 10 DCE tasks. Both C-TTO and DCE exercise were being maintained for the quantitative phase to allow different aspects of analysing preferences. The valuation sets derived by each of the valuation methods could then be compared to analyse the robustness of C-TTO and DCE in reflecting public preference.

4.4.5. Valuation outcome

Regarding the valuation outcome, some combinations of levels of attributes could be distressing to individuals who had experienced a very poor state of mental health. They tended to avoid imagining or being in a particularly low level of mental well-being. As there was only one participant who failed to reach the indifference point for some C-TTO tasks in the worse-than-dead scenario, a decision on the need to extend the amount of lead time would be investigated in the larger valuation study. Participants were asked in the quantitative phase to determine the amount of lead time required to accept life B if they failed to achieve the indifferent point even after exhausting all years in life A. Moreover, the issue of non-trading could be a potential limitation for the adoption of the C-TTO technique for the valuation of SWEMWBS due to the lack of discriminatory potential. The distribution of the derived C-TTO values would be investigated in the results of the quantitative phase, to discover any potential clustering of the values at 1. For the potential prevalence of not preferring full mental well-being, the proportion of responses with C-TTO value greater than 1 would also be investigated in the quantitative phase.

4.4.6. Overall reflections on mental well-being

In terms of practicality, the application of C-TTO and DCE valuation techniques were suitable for capturing individual attitudes towards different mental well-being scenarios.

It was encouraging that the feedback and comments received were generally positive, showing that it was feasible to gather feelings and thoughts about this valuation exercise.

It was also motivational to find that the reflection on personal mental well-being preferences, and the identification of what was important to mental health necessitated by the interviews, engaged participants in a positive way. Many seemed to enjoy the process. The interviews provide some interesting perspectives on current mental well-being literacy. Examples include some participants' belief that well-being depends on feeling useful or that optimism would not be possible without clarity of thought. Reflections on what was and wasn't valuable for mental well-being (for example the importance of challenges and the value of ups and downs), and the issues of states which were unimaginable by some participants are all of potential value to those who are seeking to improve mental well-being. This topic was beyond the scope of this thesis but the data collected might be valuable for further analysis on this subject.

The limitations of this study include its small sample size. This study was conducted as the Covid-19 pandemic was unfolding, which restricted and ultimately curtailed our ability to identify participants for face-to-face interviews. The preference data collected were highly limited to individuals within an academic environment, even though effort was exerted to include non-academic staff. The predominance of university staff or students (ten participants) in the sample may have influenced the results. The fact that data saturation was reached suggests that the study was able to identify the main issues in spite of these limitations, but there was insufficient data to assess whether the issues raised by one participant were of broader concern. Moreover, the valuation tasks were randomly allocated to participants, without tailoring tasks consistently for each participant to test the potential violation of axioms of utility theory in their responses.

4.5. Conclusion

This study constitutes the first attempt to apply health state valuation techniques to the valuation of mental well-being as measured by the SWEMWBS. The results from the cognitive interviews support the feasibility of this application and provide insights that inform the optimisation of the valuation protocol.

Chapter 5: A quantitative investigation of the feasibility, practicality and face validity of the C-TTO and DCE in the valuation of SWEMWBS

5.1. Introduction

The previous chapter documented the application of cognitive interviews to investigate the cognitive process of completing the C-TTO and DCE valuation tasks. Based on the modified mental well-being valuation protocol informed by the results of the qualitative findings, this chapter aims to quantitatively investigate the feasibility, practicality and face validity of the modified valuation protocol in a larger sample size.

5.2. Methods

Structured interviews with the presence of an interviewer (HHEY) were administered in a CAPI setting. The EQ-PVT platform (i.e. a replica of the EQ-VT 2.1) developed by the EuroQol Group (Stolk *et al.*, 2019) was used to perform the 10 C-TTO and 10 DCE tasks and record the participants' responses. Due to the COVID-19 pandemic, face-to-face interviews were not possible. All interviews were held using an online meeting software, Microsoft Teams. Individuals who were members of the general UK population, had the right to vote in the U.K. and aged 18 or above were eligible for this study. The rationale for restricting participation to UK voters was that the derived preference-based valuation set to be used in economic evaluation aims to inform democratic allocation of public sector resources in the U.K. As voters can influence societal decision making, it was valuable to understand their preferences within a mental well-being context to inform resource allocation. Participants were asked to self-declare their own voting right before participation. This research was approved by the Biomedical and Scientific Research Ethics Committee at the University of Warwick (Reference: BSREC.44/19-20).

5.2.1. Recruitment strategy

Participants were randomly drawn from a combination of convenience sampling and snowball sampling. In addition to personal networks, information of this project was advertised on social media (Facebook, Instagram, Reddit and Twitter) and other platforms including the weekly WMS Newsletter and the webpage for the CHEW to diversify the

recruitment pool and maximise the chances of recruitment. An example of the advertisement is shown in Appendix 18. Upon the completion of all interviews, 10 participants were randomly drawn. A £25 Amazon shopping voucher was given to each of the 10 winners as an expression of participation thankfulness.

5.2.2. Experimental design and sample size determination

5.2.2.1. Design for the DCE

The software Ngene was used for the generation of choice tasks. There were 50 generated choice tasks in total, divided into 5 blocks, to which respondents were randomly assigned. The number of choice tasks in each block completed by the participants was increased from 8 in the previous qualitative phase in Chapter 4 to 10 in this phase. Prior information obtained from the results of the qualitative phase was incorporated into the design to form a Bayesian D-efficient design. Appendix 19 shows the syntax for executing the design.

Additionally, a sufficient sample size was required to maintain statistical power as the valuation results were to be modelled to elicit utility values for the mental well-being states. The sample size calculation was informed by the approximate formula (Louviere *et al.*, 2000):

$$N \geq \frac{1-p}{Tp(a^2)} \left[\varphi^{-1} \left(\frac{1+\alpha}{2} \right) \right]^2 \quad \dots (3)$$

For a pairwise DCE, T (choice tasks) = 10, p (expected choice proportion) = 0.5, a (accuracy) = 0.1, α (confidence level) = 0.95, φ^{-1} = inverse normal cumulative distribution. => $N \geq 39$

As there were 5 blocks, $N \geq 39 * 5 \text{ blocks} \Rightarrow N \geq 195$. The required minimum sample size was therefore 195.

The final design generated after around 35000 evaluations performed by Ngene is provided in Appendix 20. This design possessed the lowest D-error (0.30) among all generated designs. The generated choice states did not contain highly uncommon reported states (i.e. D values of 10 or above) as identified in Appendix 17.

5.2.2.2. Design for the C-TTO

Identical to the qualitative phase, a blocked design with the use of the level-balance criterion constructed by the EuroQol group was adopted (Oppe & Van Hout, 2017). The software R was used to generate the 7 blocks of choice tasks. The number of choice tasks

in each block completed by the participants increased from 8 for the qualitative phase in Chapter 4 to 10 in this phase. The ten tasks included 2 compulsory mental well-being states [lowest mental well-being state (1111111) and one of the closer to full well-being states (4555555, 5455555, 5545555, 5554555, 5555455, 5555545 and 5555554)] plus 8 randomly generated mental well-being states. The total number of generated mental well-being states was 64. The R code required to generate this experimental design is shown in Appendix 21.

The 7 blocks generated by R are provided in Appendix 22. This selected factorial design achieved the lowest value for the level balance (27.35), compared to all other generated designs. The generated choice states did not contain highly uncommon reported states (i.e. D values of 10 or above) as identified in Appendix 17.

A summary of the C-TTO and DCE designs is provided in Table 15 below:

Table 15: A summary of the C-TTO and DCE experimental designs in the quantitative phase

| DCE design | C-TTO design |
|---|---|
| Design: a Bayesian D-efficient design | Design: A blocked design with an achieved level balance |
| Total number of choice tasks: 50 | Total number of mental well-being states: 64 |
| No. of blocks: 5 | No. of blocks: 7 |
| No. of choice tasks per participant: one block, consisting of 10 choice tasks | No. of mental well-being states per participant: one block, consisting of 10 mental well-being states [2 compulsory mental well-being states lowest mental well-being state (1111111) and one of the closer to full well-being states (4555555, 5455555, 5545555, 5554555, 5555455, 5555545 and 5555554) plus 8 randomly generated mental well-being states] |
| Sample size: ≥ 195 | |

C-TTO indicates composite time-trade off; DCE, discrete choice experiment.

5.2.3. Analysis

The robustness of the valuation protocol and the quality of data were assessed by a range of statistic indicators.

5.2.3.1. Feasibility and practicality

The cognitive burden of completing the C-TTO and DCE valuation tasks was explored. It was informed by a number of pre-designed debriefing statements that were posed at the end of the valuation exercise. Participants indicated their levels of agreement with each of the statements across a five-point Likert scale scored from 1 for “Strongly disagree”, 2 for “Somewhat disagree”, 3 for “Neither agree nor disagree”, 4 for “Somewhat agree” to 5 for “Strongly agree”. The percentage of participants indicating a particular choice for each of the statements was analysed. Participants were also asked to indicate the manageable number of tasks they were comfortable completing, and comparisons were made of the level of difficulty experienced when completing C-TTO tasks and DCE tasks. All debriefing questions are documented in Table 16.

In order to gain more insights of the application of the modified valuation protocol, numerical evidence such as the number of steps in achieving the indifference point in the C-TTO tasks recorded by the EQ-PVT, the number of respondents who failed to reach the indifference point for the C-TTO tasks, the existence of non-trading effect, and the proportion of respondents who disagreed with the rank ordering of states in the C-TTO Feedback Module were analysed.

The modelling methods of the C-TTO and DCE responses will be described in the next chapter.

Table 16: C-TTO and DCE debriefing questions

| C-TTO debriefing questions | |
|--|---|
| <u>Statement</u> | <u>Response options</u> |
| <i>I didn't have difficulty in understanding the warm-up examples of the tasks.</i> | Five-point Likert scale ranges from “Strongly disagree” to “Strongly agree” |
| <i>I didn't have difficulty in following and understanding the instructions of the tasks.</i> | Five-point Likert scale ranges from “Strongly disagree” to “Strongly agree” |
| <i>I didn't have difficulty in reaching the indifferent point between Life A and Life B for each of the trade-off tasks.</i> | Five-point Likert scale ranges from “Strongly disagree” to “Strongly agree” |

| | |
|--|---|
| <i>I didn't have technical difficulty in operating the tasks through remote control.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| <i>I didn't feel overwhelmed regarding the number of tasks (i.e. 10 trade-off tasks) required to be completed.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| DCE debriefing questions | |
| <u>Statement</u> | <u>Response options</u> |
| <i>I didn't have difficulty in following and understanding the instructions of the tasks.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| <i>I didn't have difficulty in deciding the most preferred option within each of the trade-off pairs.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| <i>I didn't have technical difficulty in operating the tasks through remote control.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| <i>I didn't feel overwhelmed regarding the number of tasks (i.e. 10 pairs) required to be completed.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| Overall debriefing questions | |
| <u>Statement</u> | <u>Response options</u> |
| <i>Overall, I didn't have difficulty in imagining each of the allocated mental well-being states.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| <i>Overall, it was manageable to complete both parts of the interview.</i> | Five-point Likert scale ranges from "Strongly disagree" to "Strongly agree" |
| <i>For the first part of the interview, were you comfortable with the number of tasks (i.e. 10 trade-off tasks) you were asked to complete?</i> | - Yes. If so, how many MORE would you have been comfortable completing? - No. If not, how many tasks would you have been comfortable completing? |
| <i>For the second part of the interview, were you comfortable with the number of tasks (i.e. 10 pairs) you were asked to complete?</i> | - Yes. If so, how many MORE would you have been comfortable completing? - No. If not, how many tasks would you have been comfortable completing? |
| <i>Overall, do you think the first part of the interview (i.e. to make time trade-off between choices of imaginable life) is cognitively easier or more difficult than the</i> | - The first part is cognitively easier than the second part. Reasons, if any: |

| | |
|---|--|
| <p><i>second part of the interview (i.e. to look at pairs of mental well-being profiles and choose the one you prefer)?</i></p> | <ul style="list-style-type: none"> - The second part is cognitively easier than the first part. Reasons, if any: - The level of difficulty for both parts of the interview is roughly the same. Reasons, if any: |
|---|--|

C-TTO indicates composite time-trade off; DCE, discrete choice experiment.

5.2.3.2. Face validity

This assessed whether the valuation techniques (C-TTO and DCE) were suitable for reflecting the preferences of participants. For the C-TTO tasks, face validity was informed by comparing the mean values of the mental well-being states. The level-sum score of the 64 selected mental well-being states was calculated. Ideally, the mean C-TTO values for the states with higher (lower) level-sum score should be higher (lower), as indication of a better (lower) mental well-being. For the DCE tasks, the level sum score for the mental well-being states in each pair was calculated. It was expected that a larger proportion of respondents would choose the option with a higher level-sum score, as it was an indication of better mental well-being.

5.2.4. Interview process

All interviews were audio and screen recorded. Potential interviewees were initially contacted by email, with positive respondents interviewed through Microsoft Teams. Instructions for installing Microsoft Teams were provided before the interviews.

- (1) The interviewer introduced the purpose of this valuation study.
- (2) The participant was asked to complete a consent form indicating their willingness to participate in this study.
- (3) The participant was introduced to the SWEMWBS descriptive system and was asked to complete the SWEMWBS in the Qualtrics software to describe own mental well-being status, followed by some background questions related to sociodemographic and socioeconomic characteristics.
- (4) The C-TTO exercise: The participant was given two versions of the warm-up example, as mentioned in the previous chapter regarding the protocol modification. The first version was identical to the warm-up example in the qualitative phase, which required imagination of mental well-being states related to a job application and relationship with

friends. The second version, which required imagination of mental well-being states related to physical health and relationship with friends, was newly added in this phase. The participant was given the flexibility to choose either one of them for practicing purposes. The participant was then guided with step-by-step explanations about the trade-off process of both better-than-death and worse-than-death scenarios, and the way of using the remote control function in Microsoft Teams to identify the preferred life on their own. The participant was also instructed that full mental well-being is defined as “all of the time” for all the seven SWEMWBS items. After the warm-up examples, three extra practice states with different levels of mental well-being were provided: high (4554545), low (2111131) and intermediate (3313432) mental well-being states. Next, the participant moved on to complete 10 tasks. After that, the rank ordering inferred by their valuations was displayed on the Feedback Module. The participant was asked to flag any disagreements or inconsistencies of the results, but they were not allowed to alter the problematic valuations. Previous research has indicated that the flagged valuations should be removed from the data (Shah *et al.*, 2014). Finally, debriefing questions around the exercise (as described in Table 16) were asked.

(5) The DCE exercise: Ten pairs of mental well-being states were presented to the interviewee and the participant was invited to choose the preferred option among two mental well-being states for each pair. The paired comparisons and the left-right order of the two mental well-being states were randomised by the EQ-PVT. Debriefing questions around the exercise (Table 16) were asked.

(6) Overall debriefing questions for both parts of the interview (Table 16) were provided to the participant, allowing the possibility to provide overall feedback and further suggestions of the tasks. It was then followed by thanking the respondent for their participation.

The participant was given the opportunity to ask any questions at each stage of the interview.

5.3. Results

All interviews were conducted between 11th December 2020 and 11th August 2021. Around 30 scheduled interviews were cancelled due to various reasons such as sickness, family emergency issues, failure to install or use Microsoft Teams on the participant’s

device, failure to ensure stable internet connection, inability to spend time for completion, and losing contact after interview confirmation, etc. In total, 227 participants attended at their scheduled date and time. However, there were two withdrawals during the completion of C-TTO exercise, due to exhaustion or anxiety, misinterpretation and misunderstanding of the tasks even after reinstructions, or technical difficulties in focusing on the task information under a screen sharing setting, etc. The total number of completed interviews was 225. The mean completion time per interview was within 60 minutes. Table 17 describes the characteristics of the 225 participants.

Table 17: Demographic characteristics of the 225 participants

| Characteristics | Counts (%) |
|--|-------------------|
| <i>Gender</i> | |
| Male | 57 (25.33) |
| Female | 165 (73.33) |
| Neutral | 1 (0.44) |
| Prefer not to say | 2 (0.89) |
| <i>Age</i> | |
| 18-30 | 34 (15.11) |
| 31-40 | 28 (12.44) |
| 41-50 | 49 (21.78) |
| 51-60 | 61 (27.11) |
| >60 | 53 (23.56) |
| Mean age | 49.16 |
| <i>Highest education level attained</i> | |
| None | 2 (0.89) |
| Grammar school | 1 (0.44) |
| GCSE | 4 (1.78) |
| O-Level | 5 (2.22) |
| A-Level | 20 (8.89) |
| Diploma | 21 (9.33) |
| Undergraduate | 71 (31.56) |
| Postgraduate | |
| Master | 69 (30.67) |

| | |
|---|-------------|
| PhD | 19 (8.44) |
| Others | 7 (3.11) |
| | |
| Others | |
| Professional qualification | 6 (2.67) |
| | |
| <i>Ethnicity</i> | |
| White | 194 (86.22) |
| Mixed / Multiple ethnic groups | 6 (2.67) |
| Asian / Asian British | 21 (9.33) |
| Black / African / Caribbean / Black British | 2 (0.89) |
| Other ethnic group | 2 (0.89) |
| | |
| <i>SWEMWBS score</i> | |
| 20 or less | 40 (17.78) |
| 21-25 | 80 (35.56) |
| 26-30 | 82 (36.44) |
| 31-35 | 23 (10.22) |
| Mean score | 24.8 |

Relative to the general UK population, these 225 participants consisted of a relatively greater proportion of females than males. The proportion of females and males in this study was 73.33% and 25.33% respectively, compared to 50.59% and 49.41% respectively in the UK (Office for National Statistics). The median age of 51 years was also higher than the median age of 40.4 years in the UK (Office for National Statistics). Even though the sample covered people with diverse education backgrounds, people with low or no qualifications were underrepresented. Nevertheless, it was promising that the ethnicity distribution of the study sample was very similar to the ethnicity distribution of the UK population, i.e. 87% of Whites, 7% of Asian/Asian British, 3% of Black/African/Caribbean/Black British, 2% of Mixed and 1% of others (Office for National Statistics).

Table 18: Response statistics of the C-TTO debriefing statements

| Item | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) | Total | Mean | Standard deviation |
|---|------------------------------|------------------------------|---------------------------------------|---------------------------|---------------------------|--------------|-------------|---------------------------|
| I didn't have difficulty in understanding the warm-up examples of the tasks. | 1 (0.44%) | 15 (6.67%) | 9 (4%) | 65 (28.89%) | 135 (60%) | 225 | 4.41 | 0.88 |
| I didn't have difficulty in following and understanding the instructions of the tasks. | 0 (0%) | 8 (3.56%) | 8 (3.56%) | 54 (24%) | 155 (68.89%) | 225 | 4.58 | 0.73 |
| I didn't have difficulty in reaching the indifferent point between Life A and Life B for each of the trade-off tasks. | 5 (2.22%) | 56 (24.89%) | 25 (11.11%) | 97 (43.11%) | 42 (18.67%) | 225 | 3.51 | 1.12 |
| I didn't have technical difficulty in operating the tasks through remote control. | 0 (0%) | 0 (0%) | 6 (3.55%) | 28 (16.57%) | 135 (79.88%) | 169* | 4.76 | 0.5 |
| I didn't feel overwhelmed regarding the number of tasks (i.e. 10 trade-off tasks) required to be completed. | 1 (0.44%) | 7 (3.11%) | 9 (4%) | 38 (16.89%) | 170 (75.56%) | 225 | 4.64 | 0.74 |

*Only 169 participants successfully used the remote-control function during the completion process.

5.3.1. Feasibility and practicality

5.3.1.1. C-TTO

Table 18 shows the response statistics of the C-TTO debriefing questions. In general, under the comprehensive step-by-step explanation of the trade-off process by the interviewer, more than 90% of participants agreed that they did not have difficulty in following and understanding the instructions of the tasks.

Regarding the warm-up examples, 98 (43.56%) participants selected the version of warm-up examples related to job application and relationship problems, whilst 127 (56.44%) participants selected the version related to physical health issues and relationship problems. This revealed the value of offering this second choice of practice example in addition to the version of job application and relationship problems in the qualitative phase. Although nearly 90% of respondents agreed that they did not have difficulty in understanding the examples, 16 participants disagreed with this statement. The high percentage of agreement might also have been partly caused by interviewer's extra clarification and explanation of participants' comprehension difficulties. The most common comprehension barrier was the information given in the examples. Some participants found that the mental feeling described within the state was incomplete and too restrictive, in the sense that the information did not reflect their whole picture of mental well-being associated with the given scenarios. For example, there was no information about the existence of external support followed by rejection of job application or treatment availability followed by health problems, as these would result in different mental reactions. Also, some participants disagreed with the mental well-being description of the example related to physical health issues. For example, imagining feeling worried followed by the confirmation of high blood pressure and diabetes was difficult, as these experiences would not necessarily be difficult experiences from their viewpoint. As a result, they were not willing to trade-off any years of life even if they were forced to imagine a feeling of anxiety. Moreover, a few participants mentioned that they found it depressing to decide between death or not.

The mean number of moves for the tasks was 4.84. It was promising that more than half (61.78%) of the participants agreed with the statement that they did not have difficulty in reaching the indifference point for the matching tasks. However, there was also a significant proportion of participants (27.11%) who disagreed with this statement,

constituting a relative low mean score (3.51) for this question when compared to other debriefing statements. The proportion of participants selecting the option “neither agree nor disagree” was also the highest (11.11%) among all other statements, revealing potential challenges of making a matching valuation decision.

Additionally, there were 56 participants who did not use the remote-control function during the process of completion, due to the reasons documented in Table 19. As the remote-control function of Microsoft Teams could only be enabled when both interviewer and interviewee were using the Teams app on a computer/laptop, it could not be used when the participant was using the app in a tablet (15 participants), the Teams mobile app in a phone (16 participants), or the web browser version of Teams (21 participants). Also, a few participants also failed to use the remote-control function even if the function was enabled as they did not possess the necessary computer skills to move around the cursor to the appropriate area and click the desired buttons by themselves. There was also one rare case where the clicks detected by Teams did not match with what the participant actually clicked. In these cases, the interviewer clicked the buttons on behalf of the participants.

Table 19: Reasons for not using the remote-control function

| Reasons | Number of participants |
|--|-------------------------------|
| Tablet | 15 |
| Phone | 16 |
| Web browser | 21 |
| Bad computer skills (burden)/ other technical issues | 4 |

These 56 responses for the debriefing question regarding the difficulty in remote-control operation were deleted, resulting in 169 remaining responses for those who used the remote-control function. It was discovered that more than 95% of these participants agreed that they did not have technical difficulty in using the remote control, once the remote-control function was successfully implemented. No participant expressed “somewhat disagree” or “strongly disagree” to this statement. The mean score for this statement was also the highest.

Furthermore, more than 90% of the participants did not feel overwhelmed regarding the number of tasks (i.e. 10 tasks) required to be completed. A few of them suggested the elimination of the length or the number of practice questions prior to the completion of 10 tasks. Simpler practices with less information digestion could help increase their engagement in completing the 10 tasks.

Apart from the support of feasibility or practicality based on the C-TTO debriefing questions, the responses to the 10 trade-off tasks were analysed. Overall, 13 participants (5.78%) failed to reach the indifference point for particular states. Among them, 10 failed to reach the indifference point even after exhausting all extra 10 years of lead time in the worse-than-death scenario. The cases of these 10 participants are described in Table 20:

Table 20: Failure to reach the indifference point for the worse-than-death scenario

| Participant | Number of tasks | Number of extra years of lead-time to reach the indifference point | State | Implied TTO value |
|-----------------|-----------------|---|--|-------------------|
| 1 st | 1 | 20 | 1233323 | -3 |
| 2 nd | 1 | Won't reach the indifference point, no matter how many extra years were provided. | 1111111 | $-\infty$ |
| 3 rd | 8 | 10 | 1111111, 1415144, 3221211, 1322113, 4333525, 3531142, 5123554, 2444432 | -2 |
| 4 th | 1 | 2 | 2521112 | -1.2 |
| 5 th | 2 | Won't reach the indifference point, no matter how many extra years were provided. | 1111111, 1112211 | $-\infty$ |
| 6 th | 3 | 5 | 3221211, 1111111, 1322113 | -1.5 |
| 7 th | 3 | 10 | 5142251, 3243413, 4112255 | -2 |
| 8 th | 1 | Won't reach the indifference point, no matter how many extra years were provided. | 1111111 | $-\infty$ |

| | | | | |
|------------------|---|---|--|-----------|
| 9 th | 3 | Won't reach the indifference point, no matter how many extra years were provided. | 1111111, 5211141, 5211424 | $-\infty$ |
| 10 th | 2 | 5 | 1412452, 1112211 | -1.5 |
| | 6 | Won't reach the indifference point, no matter how many extra years were provided. | 3522134, 2554521, 3243413, 5142251, 4112255, 1111111 | $-\infty$ |

As each of the 225 participants completed 10 tasks, there were 2250 C-TTO responses in total. The total number of tasks for which the participants failed to achieve the indifference point in the worse-than-death scenario was 31, occupying only 1.38% of the total responses. Five participants declared that they failed to find the amount of full mental well-being years equivalent to some states, no matter how many extra years of lead-time given. These implied a C-TTO value of $-\infty$. The lowest mental well-being state 1111111 was commonly valued as $-\infty$ for these five participants.

Furthermore, as described in Table 21 below, 3 participants failed to reach the indifference point for some states in the better-than-death scenario:

Table 21: Failure to reach the indifference point for the better-than-death scenario

| Participant | Number of tasks | State | Implied TTO value |
|-----------------|-----------------|--|-------------------|
| 1 st | 1 | 5545555 | >1 |
| 2 nd | 10 | 4255431, 1441325, 4555555, 3115413, 1111111, 1151143, 4512541, 2323241, 5124354, 2521214 | >1 |
| 3 rd | 6 | 1433453, 4121554, 4352232, 2432335, 5555455, 2145542 | >1 |

The total number of tasks for which the participants failed to achieve the indifference point in the better-than-death scenario was 17, occupying only 0.76% of the total responses. Full mental well-being was undesirable for them when there was no room for life improvement. All these responses implied a TTO value of >1. The C-TTO tasks failed to proceed when

they preferred life B (i.e. a state lower than full mental well-being) rather than life A (i.e. full mental well-being) at the beginning of the task. It was interesting to know that one participant exhibited a strict rejection of the idea of full mental well-being as the implied values for all 10 C-TTO tasks were >1 . The lowest mental well-being state 1111111 was even better than full mental well-being for this participant.

Concerning the Feedback Module, 151 (67.11%) out of 225 participants did not indicate any disagreement with the rank ordering of their 10 completed C-TTO tasks. Among the 74 participants (32.89%) who indicated disagreement with particular states, 28 participants flagged disagreement with 1 state, 27 participants flagged disagreement with 2 states, 12 participants flagged disagreement with 3 states, 6 participants flagged disagreement with 4 states, and 1 participant flagged disagreement with 5 states. After deleting the 147 flagged answers, 2103 valid responses remained.

In addition, non-trading effects for the C-TTO responses were also investigated. After deleting the flagged responses, 154 participants were not willing to trade off any years of full mental well-being (i.e. C-TTO value of 1) for at least one task. For these 154 participants, 104 participants were not willing to trade off any years of full mental well-being for one task, 26 participants for two tasks, 12 for three tasks, 6 participants for four tasks, 1 participant for five task, 2 participants for six tasks, and 1 participant for seven tasks. One participant experienced a close to perfect non-trading effect (i.e. not willing to trade off years of life for all states except the state 1111111). Remarkably, one participant experienced a perfect non-trading effect (i.e. not willing to trade off years of life for all 10 C-TTO states). The implied TTO values for all 10 tasks were 1 in this case.

Table 22: Response statistics of the DCE debriefing statements

| Statement | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) | Total | Mean | Standard deviation |
|--|------------------------------|------------------------------|---------------------------------------|---------------------------|---------------------------|--------------|-------------|---------------------------|
| I didn't have difficulty in following and understanding the instructions of the tasks. | 0 (0%) | 1 (0.44%) | 1 (0.44%) | 22 (9.78%) | 201 (89.33%) | 225 | 4.88 | 0.38 |
| I didn't have difficulty in deciding the most preferred option within each of the trade-off pairs. | 27 (12%) | 76 (33.78%) | 22 (9.78%) | 67 (29.78%) | 33 (14.67%) | 225 | 3.01 | 1.31 |
| I didn't have technical difficulty in operating the tasks through remote control. | 0 (0%) | 1 (0.59%) | 7 (4.14%) | 22 (13.02%) | 139 (82.25%) | 169* | 4.77 | 0.55 |
| I didn't feel overwhelmed regarding the number of tasks (i.e. 10 pairs) required to be completed. | 0 (0%) | 9 (4%) | 10 (4.44%) | 36 (16%) | 170 (75.56%) | 225 | 4.63 | 0.75 |

*Only 169 participants successfully used the remote-control function during the completion process.

5.3.1.2. DCE

Table 22 shows the response statistics for the DCE debriefing questions. Compared to the C-TTO, the instruction for the completion of DCE was much shorter, as the participants were only instructed to click on the preferred option between states A and B. Nearly 90% of the participants selected the response option “strongly agree” to the statement “I didn’t have difficulty in following and understanding the instructions of the tasks.” Only 1 participant selected “somewhat disagree” to this statement. In line with expectations, the mean score for this statement (4.88) was the highest across the relevant statements, indicating no instruction comprehension issues for the participants in general.

Although the instruction of tasks was easier to understand, the decision of the preferred option for each pair was cognitively challenging to some participants. Less than half of the participants (44.45%) somewhat agreed or strongly agreed with the statement that they didn’t have difficulty in selecting the most preferred option within each of the trade-off pairs. The proportion of participants who somewhat disagreed and strongly disagreed with this statement were 33.78% and 12% respectively. The mean score (3.01) was also the lowest, revealing a mixed response to this statement.

As mentioned before, only 169 participants successfully used the remote-control function. The response statistics to the statement regarding the technical issue of using remote-control function was promising, with more than 80% of participants who strongly agreed that they did not have technical difficulty in the remote-control operation. the participants’ speed of completing the DCE tasks was sometimes affected by the remote-control sensitivity. The remote-control function in Microsoft Teams became inactive when there was no click or cursor moving after around 3-5 seconds. As a result, before they could successfully click their preferred option for each task based on their decision in mind, participants usually required constant reactivation of the remote-control function by continuous and large extent cursor moving.

Moreover, more than 90% of the participants agreed that they did not feel overwhelmed regarding the number of tasks (i.e. 10 pairs) required to be completed. This suggested that the difficulty in making trade-off decision for participants did not make them less manageable in terms of the number of tasks to be completed.

In addition, when it came to the DCE responses, there were 4 missing answers due to technical issues of the EQ-PVT. The EQ-PVT platform failed to display three tasks to one participant and one task to another participant, resulting in the total completion of only seven tasks for the former participant and nine tasks for the latter participant. As each of the 225 participants completed 10 tasks, the number of valid responses was reduced from 2250 to 2246. Also, potential strategic pattern or sign of lacking engagement was discovered for the DCE responses of four participants (BBBAABBBAA, BBBBAAAAA, AAAAAAAAAA, ABAABABAAB). However, these responses were not deleted to avoid the introduction of sample selection bias. As the interviewer was monitoring the whole process of the interview in-person under a virtual face-to-face setting, all interviewees maintained certain level of engagement and concentration during the task completion. It was difficult to justify that their answers were random and invalid.

Table 23: Response statistics of the overall debriefing statements

| Item | Strongly disagree (1) | Somewhat disagree (2) | Neither agree nor disagree (3) | Somewhat agree (4) | Strongly agree (5) | Total | Mean | Standard deviation |
|--|------------------------------|------------------------------|---------------------------------------|---------------------------|---------------------------|--------------|-------------|---------------------------|
| Overall, I didn't have difficulty in imagining each of the allocated mental well-being states. | 4 (1.78%) | 29 (12.89%) | 18 (8%) | 89 (39.56%) | 85 (37.78%) | 225 | 3.99 | 1.07 |
| Overall, it was manageable to complete both parts of the interview. | 0 (0%) | 0 (0%) | 2 (0.89%) | 40 (17.78%) | 183 (81.33%) | 225 | 4.8 | 0.42 |

5.3.1.3. Overall impression

Each participant was given some overall debriefing questions covering both the first part (i.e. C-TTO) and the second part (i.e. DCE) of the interview. Table 23 shows the response statistics of the two statements about the imagination burden and manageability of the tasks.

Around 80% of the participants agreed that they did not have difficulty in imagining the hypothetical mental well-being states. However, there were also more than 10% of the participants who struggled to understand the hypothetical scenarios. Nevertheless, it was encouraging that 99.11% of the participants agreed with the statement “Overall, it was manageable to complete both parts of the interview”, with the mean score of 4.8. No participant selected “somewhat disagree” or “strongly disagree” to this statement.

Upon providing responses to these two statements, participants were asked the manageable number of tasks for both parts of the interview. For both the C-TTO and DCE, the results were roughly the same, in the sense that 215 participants found it manageable to complete the 10 given tasks or even more. Only 10 participants thought that they were not comfortable with the number of tasks and were comfortable to complete less than 10 tasks. The mean manageable number of tasks for the C-TTO and DCE was 14.91 (SD: 7.98) and 15.29 (SD: 8.16) respectively. This revealed that, in addition to the 10 given tasks for each part, participants were comfortable to complete around 5 additional tasks on average.

Finally, participants were asked to compare the level of difficulty associated with C-TTO and DCE tasks. In general, compared to the C-TTO tasks, the results showed that the DCE tasks were slightly less challenging to participants. There were 94 (41.78%) participants who thought that the DCE tasks were cognitively easier than the C-TTO tasks, whereas 75 (33.33%) participants thought that the C-TTO tasks were cognitively easier. Notably, 56 (24.89%) participants thought that the level of difficulty between both parts was roughly the same.

5.3.2. Face validity

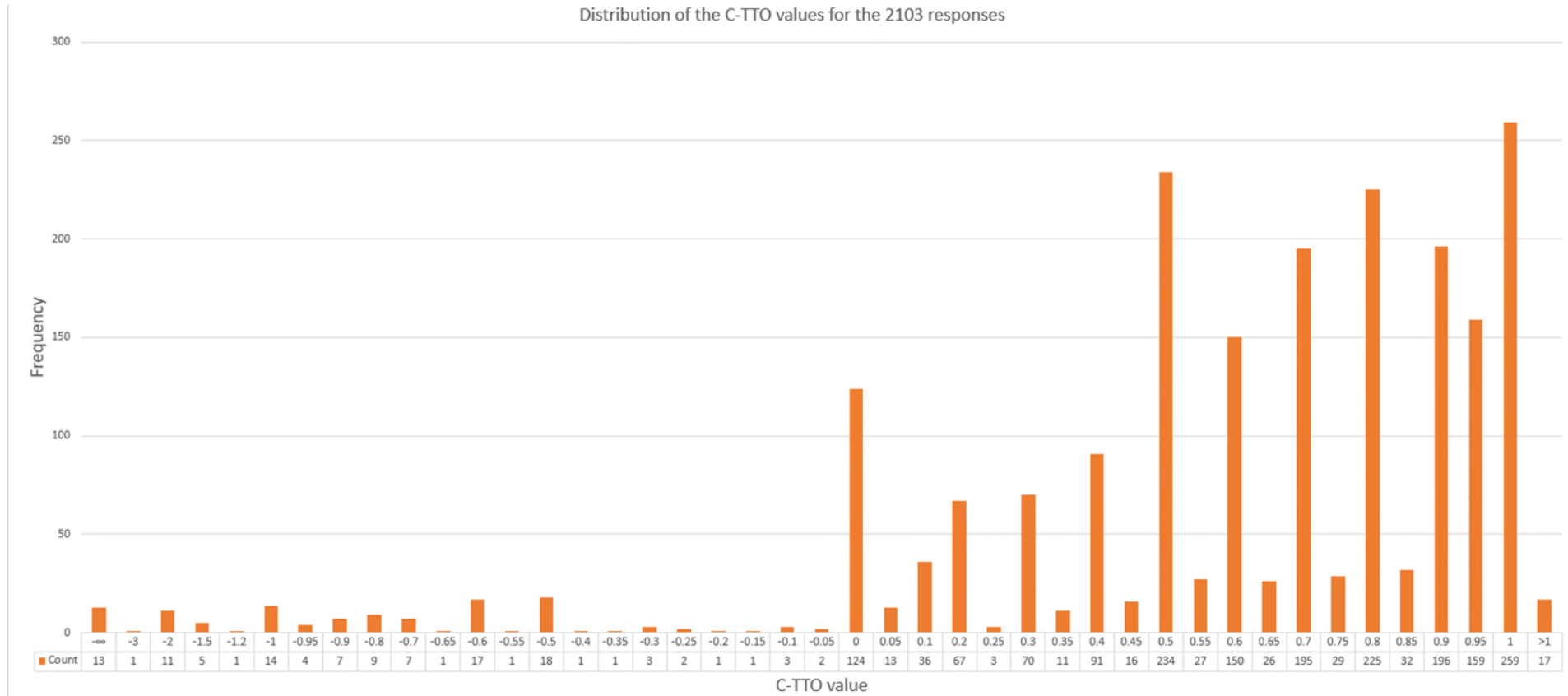
5.3.2.1. C-TTO

As mentioned before, 147 flagged responses in the Feedback Module were deleted, resulting in 2103 valid responses. Figure 7 below shows the distribution of the C-TTO values of the 2103 responses.

The distribution was left-skewed, in the sense that more C-TTO values were clustered at the positive end. The most common C-TTO value was 1, as indicated by the peak of the distribution. The second most frequent value was 0.5, followed by the value of 0.8. There were 123 responses with negative C-TTO values, which contributed only 5.85% of the total responses.

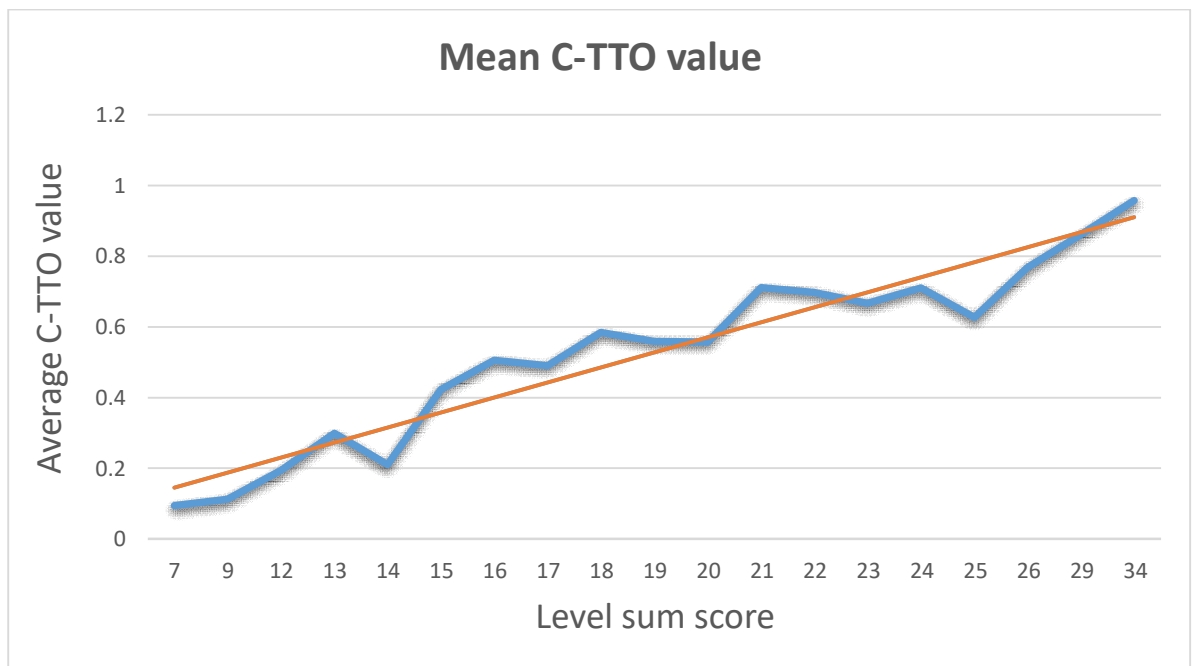
A graphical plot of the mean C-TTO value against the level-sum score of the 64 selected mental well-being states, as shown in Figure 8 below, supported the face validity of the C-TTO technique. The overall trend showed that states with higher level-sum score were generally associated with higher mean C-TTO values.

Figure 7: Distribution of the C-TTO values



C-TTO indicates composite time-trade off.

Figure 8: Relationship between mean C-TTO value and level-sum score



Notes: The responses $-\infty$ and >1 were not included in the calculation of mean values.

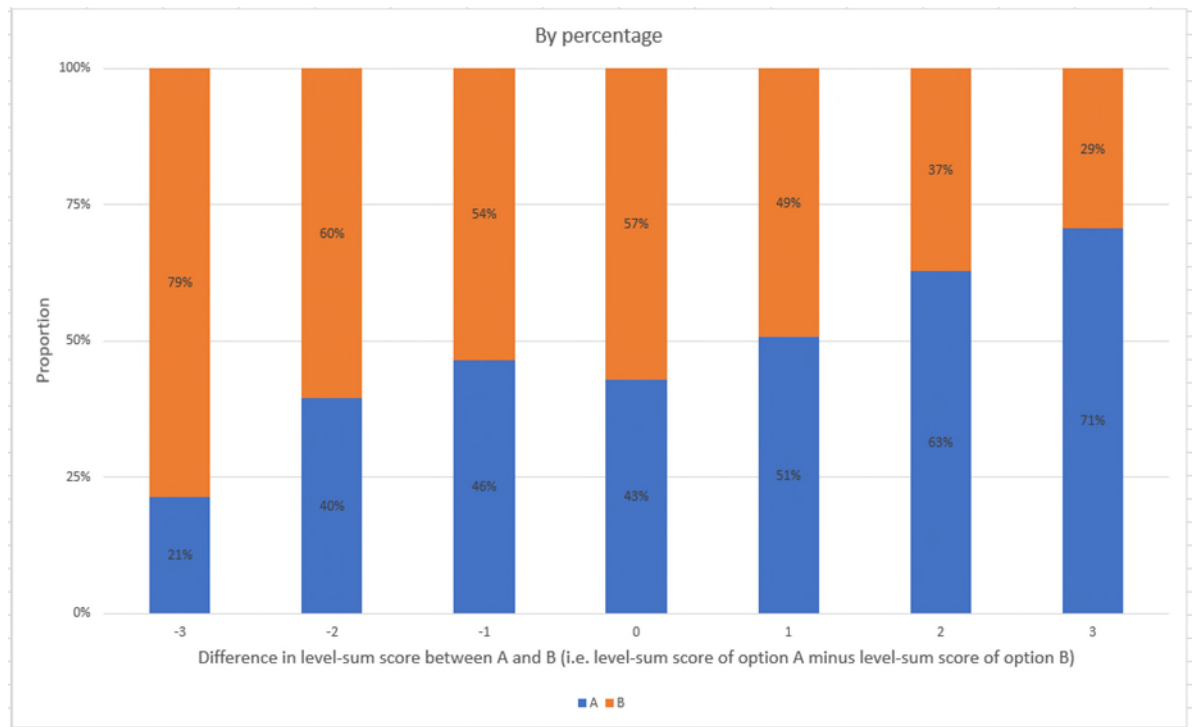
C-TTO indicates composite time-trade off.

5.3.2.2. DCE

Figure 9 shows the relationship between the percentage of participants and the difference in level-sum score between A and B, which was calculated by the level-sum score of option A minus the level-sum score of option B. Specifically,

$$\text{Difference in level sum score} = \begin{cases} < 0, & \text{if } A < B \\ 0, & \text{if } A = B \\ > 0, & \text{if } A > B \end{cases}$$

Figure 9: Relationship between the percentage of the chosen option and the difference in level-sum score among the 2246 responses



| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---------------------|-------------|--------------|--------------|--------------|--------------|--------------|-------------|
| A was chosen | 19 (21%) | 72 (40%) | 310 (46%) | 213 (43%) | 228 (51%) | 142 (63%) | 96 (71%) |
| B was chosen | 70 (79%) | 110 (60%) | 357 (54%) | 284 (57%) | 221 (49%) | 84 (37%) | 40 (29%) |
| Count | 89 | 182 | 667 | 497 | 449 | 226 | 136 |

It was expected that participants preferred option A (B) when the difference in level-sum score was positive (negative), as the level of mental well-being in state A was higher (lower) than level of mental well-being in state B. The results generally supported the face validity of the DCE technique. It was evident that a majority of participants preferred option B at the lower-end of the difference in level-sum score at -3 (Option A: 21% v.s. Option B: 79%). The proportion of participants who chose option B diminished when the difference in level-sum score was smaller. On the other hand, the proportion of participants who chose option A slightly dominated that of option B when the difference in level-sum score was positive at 1 (Option A: 51% v.s. Option B: 49%). The proportion of participants selecting option A increased with an increasing difference in level-sum

score. The gap between the selection of option A and option B was the largest when the difference in level-sum score was 3 (Option A: 71% v.s. Option B: 29%).

5.4. Discussion

This chapter explored the feasibility, practicality and face validity of the modified valuation protocol based on the interview responses from the 225 participants. Different from other pilot studies that tested the feasibility, practicality and face validity of the C-TTO and DCE techniques (Janssen *et al.*, 2013; Papadimitropoulos *et al.*, 2015), this study was the first attempt to use Microsoft Teams as a means of conducting virtual face-to-face interviews for the completion of C-TTO and DCE tasks. Generally speaking, the results of the debriefing questions and the statistical analyses of the C-TTO and DCE responses supported the feasibility and practicality of the modified valuation protocol.

Firstly, unlike the DCE tasks, the procedure of completing the C-TTO tasks included more technical steps, as the adjustment of life years within a timeline with the position of death was required (Louviere & Woodworth, 1983; Lugnér & Krabbe, 2020). Even though it was time consuming to explain the C-TTO concepts to participants, previous research around the adoption of the C-TTO under the EQ-VT protocol emphasised the important role of a professional interviewer in communicating the multistep procedure to participants (Oppe *et al.*, 2016; Yang *et al.*, 2017). Too short a time spent on explaining the practice tasks and no explanation of the worse-than-death scenario lowered the quality of data collected (Stolk *et al.*, 2019). Despite the concern about cognitive understanding, it was encouraging that participants in general did not have difficulty in comprehending both the C-TTO and DCE instructions provided by the interviewer in this study, as indicated by the consistently high mean scores for the instruction comprehension statement for both parts of the interview. This implied that the tedious C-TTO process was not necessarily a negative experience for participants.

Secondly, even though most participants did not have difficulty in understanding the C-TTO warm-up examples, careful thought was required to decide on the information included in the scenario description. Participants sometimes argued about the insufficient or incorrect mental well-being information described in the job application or physical health examples. Much time was taken to clarify the meaning of scenarios as different participants could have different mental interpretations associated with the given scenarios.

The confusion of information given in the C-TTO practice examples was not documented in the wheelchair examples used in the EQ-5D-5L valuation studies (Ramos-Goni *et al.*, 2017a). The reason for lacking controversial elements within the wheelchair examples could be due to the main focus on physical dimensions within the EQ-5D-5L. It was a clear fact that being in a wheelchair was a discount of full health (i.e. perfect health) in terms of mobility, self-care and usual activities, etc. This could be explained by the generic nature of physical health issues, as these were applicable to people with different ages. However, instead when participants were asked to imagine the effect of physical health problems or being rejected for job application on the level of mental well-being, it was debatable as various external factors could influence the mental feeling. In this sense, it was understandable that some participants disagreed with the information given in the practice scenarios of this study. The discount to full mental well-being followed by physical health problems and job application rejection might be uncertain and could not be captured within a few sentences. Moreover, there was no comprehension issue identified in the last three practice states regarding the valuation of three SWEMWBS states: 4554545, 2111131 and 3313432. It was not surprising as SWEMWBS is already a robust measure which sufficiently captured multifaceted aspects of mental well-being attributes. There was a clear distinction between these three practice states and full mental well-being (555555) in life A. To avoid unnecessary interpretation confusion, future SWEMWBS valuation exercises could consider replacing the first three practices related to job applications and relationship problems, or physical health issues and relationship problems by other value-added exercises. For example, participants could be asked to familiarise themselves with the SWEMWBS by ranking the importance of the seven SWEMWBS items. They could then have an idea of their own weight assigned to each attribute before rolling out to the valuation tasks. Also, some practice tasks for them to play around the timeline could be included as participants were sometimes confused about the iteration algorithm, as each move could mean an upward or downward titration with an increment of either 1 year or 6 months after the first three steps of each task (Oppe *et al.*, 2016).

Furthermore, it was interesting that participants in general found it more difficult in deciding the preferred option within each DCE pair than deciding the indifference point for each of the C-TTO tasks. However, the proportion of participants who thought that their overall experience of completing the choice-based DCE tasks were cognitively easier

slightly surpassed the proportion to that of the matching-based C-TTO tasks, when they were asked to compare the level of difficulty between both parts of the interview. It seems that the DCE exercise was superior to the C-TTO exercise in terms of its simple and time efficient procedure. Also, it should be noted that the difficulty in deciding the preferred DCE option was not necessary a downside of this valuation technique. It could mean that the purpose of “trade-off” was better achieved when participants spent more cognitive effort in trading off between items and levels. In this sense, the DCE tasks could be a time-efficient valuation technique and the responses could be reliable in reflecting the trade-off information during the completion process.

In addition, participants were comfortable in using the remote-control function in Microsoft Teams during the completion of C-TTO and DCE tasks. However, there were still many hurdles in implementing the remote-control successfully. The remote-control function was still restrictive and immature in Microsoft Teams as it could only be enabled when the participant was using the Teams app in a laptop or a computer. It was also not sensitive enough to detect participants’ action, as participants were often required to move around the cursor to activate the remote-control function. It was true that these limitations could be resolved by allowing interviewer to click every button on behalf of participants. However, the original idea of enabling remote-control function was to allow participants to complete the tasks on their own without disturbance from the interviewer, which might cause anxiety. There were other technical problems identified when conducting the interviews through Microsoft Teams. Even though the installation instruction of Microsoft Teams was provided to each participant, participants sometimes reported download or login problems related to authentication, and connection problem of the app, etc. Using web browser version of the Teams could be a solution to the problems discovered in the app version, as it did not require the installation of app and users were not required to create a Teams account. There would be no login issue in this sense. However, the use of web browser version came at a cost. In addition to its inability to enable the remote-control function, it performed poorly in the screen sharing function due to a constant delay of the contents displayed on the participant’s screen. The problem of screen freezing was also recognised. As a result, sometimes there was a mismatch between the contents displayed on the interviewer’s screen and the contents displayed on the participant’s screen. This caused confusion to participants as the audio instruction of task completion might not correctly describe the contents participants were looking at. In terms of other meeting

software, I could not test the feasibility of Zoom, due to its prohibited use in research interview under the data security policy at the University of Warwick.

Regarding the nature of tasks, previous literature has documented concerns around using the C-TTO as a valuation technique, such as the existence of non-traders and the failure to achieve the indifference point in the worse-than-death scenario (Attema *et al.*, 2013). These issues were also discovered in this research, but the proportion was small. There were only six participants with non-trading effect for at least half of the 10 C-TTO tasks, representing a tiny proportion (2.67%) of all participants. Also, only 5.78% of the participants failed to reach the indifference point for particular states. Interestingly, similar to the results of the qualitative phase, there was an observation of C-TTO value greater than 1, which was not discovered in other studies that applied the C-TTO valuation technique. This non-monotonic preference indicated that the highest level of mental well-being was not necessarily the most preferred status for participants, as participants sometimes pursued diverse levels of mental well-being in their life, i.e. full mental well-being was not always the healthiest mental status. All these practical concerns were not discovered in the completion of DCE tasks. Also, no participants failed to choose the preferred option within each DCE pair, revealing that the response setting of DCE tasks was less controversial than that of the C-TTO in practice.

Concerning the optimal number of tasks, most participants declared their ability to complete more C-TTO and DCE tasks. Future SWEMWBS valuation study could increase the number of tasks for each participant, taking into account the potential deterioration in concentration level when the mean interview time increases further beyond 60 minutes.

Apart from that, similar to the distribution of the C-TTO responses in the England EQ-5D-5L valuation study (Devlin *et al.*, 2018), clustering at the values 0, 0.5 and 1 was also found in this study, but to different extents. The most frequent value in the EQ-5D-5L valuation study was 0.5, followed by 0 as the second frequent value, and 1 as the third. For this study, the most frequent value was 1, followed by 0.5 as the second, and 0.8 as the third. The value of 0 was ranked as the 8th frequent value. Nevertheless, the face validity of C-TTO and DCE valuation techniques was confirmed in this study. The mean C-TTO value was in general higher for states with higher level-sum score, and vice versa. Participants also tended to choose a state with higher level-sum score within each DCE pair. It should be acknowledged that a good face validity of the C-TTO task was partly

due to the improved quality control of the EQ-VT protocol. The Feedback Module of the C-TTO was firstly introduced in the EQ-VT 2.0 (Stolk *et al.*, 2019). It was then reserved in the EQ-VT 2.1 adopted in this study. The purpose of this Feedback Module was to allow participants to indicate any disagreement with the rank ordering of the 10 C-TTO tasks inferred by their valuations. This Feedback Module was useful in detecting illogical C-TTO values of states, as participants might not be consistent in their standard of finding the indifference point for every state. More than 30% of the participants utilised this Feedback Module to flag certain states with incorrect rank ordering. The deletion of these flagged states indeed improved the quality of the collected C-TTO data, in the sense that states with illogical values were eliminated. The real preference of participants was better reflected in the remaining data.

There were limitations to the result presented in this chapter, due to the constraints on sample recruitment and interview protocol. The first one was the concern on sample diversification. The demographic diversity was limited when convenience sampling and snowballing were the main recruitment strategies. The demographic distribution of the participants in this study was not truly representative of the general UK population, even though effort was made to purposely diversify the recruitment pool under limited time and resources. Facebook was the main media for advertising the interview recruitment of this study, as it was time efficient in terms of having the potential to reach more than 600 members of the general population per £1 spent in a day. The demographic target of a specific country with different age groups can also be easily tailored. Also, it covered a wide variety of advertising placements, including mobile app news feed, Instant Article, Instagram Stories, Instagram feed, and Mobile in-stream video, etc. However, there were difficulties in sampling a balanced level of men versus women. According to the Facebook advertising statistics in Appendix 23, men were also found to have a much lower approach rate relative to women. The interview recruitment was advertised in Facebook for 23 times, each with specific duration and cost. Among these 23 occasions, 21 of them were targeted at the recruitment of both men and women. The women approach rate for each time was higher than that of the men, with the proportion difference ranged from a minimum of 22.40% to a maximum of 90.20%. The cost per link click to the application form of this interview ranged from a minimum of £0.09 to a maximum of £0.26. Remarkably, among the 2 times with a target at the recruitment of men only, the cost per link click was £0.32 and £0.41 respectively, which was the highest across all 23 times of advertisement. This

implied that it was difficult and costly to approach men. Furthermore, there was a diminishing return in approaching diverse members of the population, as the advertisement might appear to the same individual several times. It would be costly to find more participants in this sense. Moreover, there was no option in Facebook to purposely target at specific group of members with certain education background, contributing to the difficulty in obtaining participants with relatively low qualifications. Another limitation of this study was that only one interviewer was responsible for all interviews. The protocol compliance performance of the interviewer was not constantly monitored by a team of trained colleagues. However, sufficient rehearsals were ensured before rolling out to the interviews with participants. Also, the advantage of having one interviewer is that the interview protocol was consistent and there was an absence of interviewer effect between interviewers, which was commonly found in other large-scale valuation studies (Attema *et al.*, 2013; Devlin *et al.*, 2018; Lugnér & Krabbe, 2020). Moreover, unlike other valuation studies which strictly followed the quality control criteria of the EQ-VT (Ramos-Goni *et al.*, 2017a), the time spent by interviewer on explaining the C-TTO practice tasks and the time spent by participants in completing the C-TTO tasks were not analysed in this study. The data of time spent on the task was not meaningful as it was greatly influenced by the internet stability between interviewer and interviewee, and the ability of participants to use the remote-control function. Also, as the interviewer was more adapted to the remote-control function, the time of completion was usually shorter when interviewer clicked every button on behalf of the participants due to the failure to enable the remote-control function. It was also argued that a short completion time did not necessarily imply the potential short-cutting approach or laziness of participants, given that different participants could have different speeds of processing the trade-off information. Lastly, considering the government rules in household mixing and health safety of the interviewer under the Covid-19 pandemic, all interviews were conducted virtually online. The participation eligibility was then determined by the internet accessibility. A number of interviews by participants was missed as they did not have electronic device or internet connection to support the interviews in Microsoft Teams. Future research could investigate the feasibility of combining both in-person face-to-face and online face-to-face modes of interview.

5.5. Conclusion

This study found that the use of C-TTO and DCE in the valuation of SWEMWBS based on the modified valuation protocol from the qualitative phase was feasible or practical under a virtual face-to-face interview setting. The responses to the valuation tasks also confirmed the face validity of the C-TTO and DCE valuation techniques. The result of this study has the potential to inform future SWEMWBS valuation studies for the development of a national UK valuation set.

Chapter 6: Modelling preliminary versions of preference-based valuation set

6.1. Introduction

Chapter 5 presents work that confirmed the feasibility, practicality and face validity of the C-TTO and DCE valuation techniques for the valuation of the SWEMWBS. With a view to understanding the relative importance of each attribute level of the SWEMWBS, this chapter aims to model and compare preliminary preference-based valuation sets of the SWEMWBS based on the C-TTO and DCE responses from the structured interviews.

6.2. Methods

Econometric modelling techniques were adopted to estimate preference scores for all 78,125 mental well-being states generated by the SWEMWBS. Stata 16.0 was used to generate all modelling outputs.

6.2.1. Heteroskedastic Tobit model for the C-TTO data

The C-TTO task consisted of two parts: conventional TTO for the valuation of better-than-dead states and lead-time TTO for the valuation of worse-than-dead states. As the number of years of full mental well-being in life A lay between 0 and 10 and the number of years of the lower than full-mental well-being state in life B was set at 10 years in the better-than-dead scenario, the generated value was bounded between 0 and 1. The value generated from the worse-than-dead scenario lay between -1 and 0, given that the ratio of lead-time to duration of the mental well-being state was 1:1. The C-TTO value generated by both better-than-dead and worse-than-dead scenarios therefore ranged between -1 and 1, assuming that the participant was able to reach the indifference point between life A and life B. As mentioned in the previous chapter, there were observations of values lower than -1 and greater than 1 in this study due to participants' failure to achieve the indifference point for some tasks. However, the proportion of these responses was small, contributing to only 2.28% out of the total number of 2103 responses after excluding the values of flagged states in the Feedback Module. In this sense, a limited dependent variable model could be used to censor these values to the interval of [-1,1]. It should be noted that the ordinary least squares regression was not suitable for modelling the censored sample, as the generated estimators would be biased and inconsistent (Heij *et al.*, 2004; Long & Long, 1997; Wooldridge, 2015). Instead, a Tobit model with the dependent

variable left censored at -1 and right censored at 1 was used to model the C-TTO data. The econometric specification of the main effects regression was as follows:

$$Y_{C-TTO} = \beta_0 + f(\text{ITEM}_{iLj}\beta) + \varepsilon \quad \dots (4)$$

$$\begin{aligned} Y_{CTTO} = & \beta_0 + \beta_1\text{ITEM1}_{L1} + \beta_2\text{ITEM1}_{L2} + \beta_3\text{ITEM1}_{L3} + \beta_4\text{ITEM1}_{L4} + \beta_5\text{ITEM2}_{L1} \\ & + \beta_6\text{ITEM2}_{L2} + \beta_7\text{ITEM2}_{L3} + \beta_8\text{ITEM2}_{L4} + \beta_9\text{ITEM3}_{L1} + \beta_{10}\text{ITEM3}_{L2} \\ & + \beta_{11}\text{ITEM3}_{L3} + \beta_{12}\text{ITEM3}_{L4} + \beta_{13}\text{ITEM4}_{L1} + \beta_{14}\text{ITEM4}_{L2} + \beta_{15}\text{ITEM4}_{L3} \\ & + \beta_{16}\text{ITEM4}_{L4} + \beta_{17}\text{ITEM5}_{L1} + \beta_{18}\text{ITEM5}_{L2} + \beta_{19}\text{ITEM5}_{L3} + \beta_{20}\text{ITEM5}_{L4} \\ & + \beta_{21}\text{ITEM6}_{L1} + \beta_{22}\text{ITEM6}_{L2} + \beta_{23}\text{ITEM6}_{L3} + \beta_{24}\text{ITEM6}_{L4} \\ & + \beta_{25}\text{ITEM7}_{L1} + \beta_{26}\text{ITEM7}_{L2} + \beta_{27}\text{ITEM7}_{L3} + \beta_{28}\text{ITEM7}_{L4} \\ & + \varepsilon \end{aligned}$$

where

Y_{CTTO} = A dependent variable for the model, representing the C-TTO value

β_0 = A constant term of the model.

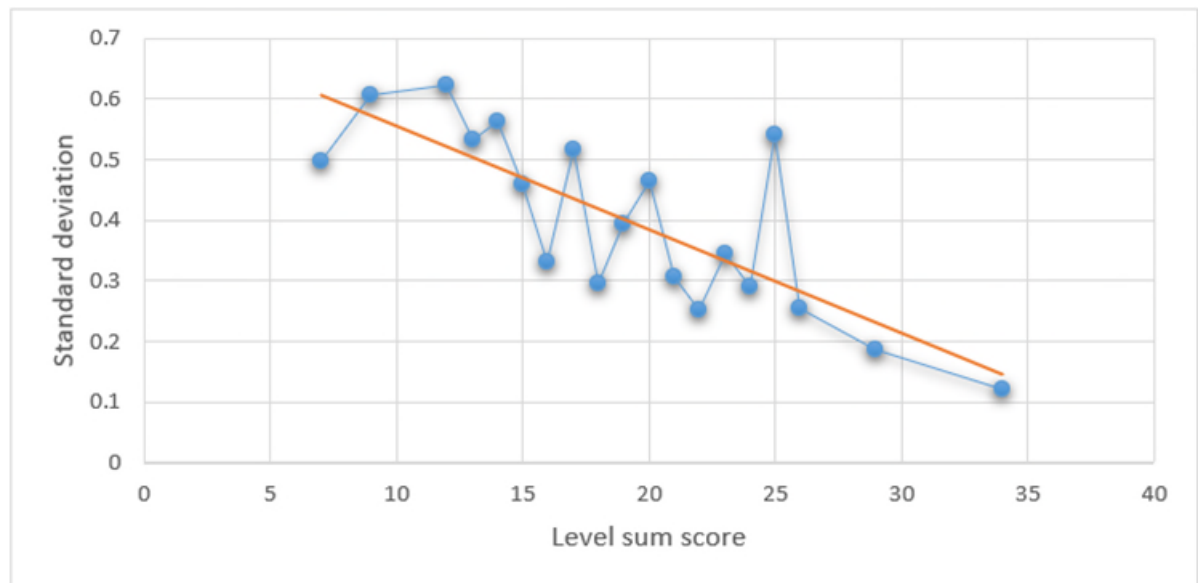
ITEM_{iLj} = A group of binary dummy variables ITEM_i for a specific level L_j ($i = 1,2,3, \dots,7$ and $j = 1,2,3,4$).

β = A group of coefficients for the explanatory variables indicating utility changes from the highest mental well-being state.

ε = An error term of the model.

As SWEMWBS is a seven-item scale with five levels for each of the seven items, there was an incorporation of 28 dummy variables (four dummies for each of the seven items) within the regression equation, with level five (the highest mental well-being level) as the reference category for each item. Also, as the standard deviation of the C-TTO response data generally increased with the lower mental well-being state (Figure 10), there was an obvious violation of the homoskedastic assumption for the residuals. To accommodate the heteroskedastic nature of the explanatory variables, a Heteroskedastic Tobit model was used, assuming that the residuals were normally distributed.

Figure 10: Relationship between the standard deviation of the C-TTO responses against the level-sum score



Notes: The responses $-\infty$ and >1 were not included in the calculation of standard deviation.

Moreover, results from the think-aloud process of the qualitative phase (Chapter 4) suggested that a participant with the experience of mental illness tended to give lower C-TTO values to mental well-being states, compared to those without indication of personal mental distress. To explore the relationship between mental well-being status and C-TTO values, the SWEMWBS score obtained by participants' completion of the SWEMWBS was added as an individual-specific covariate to the model. Furthermore, a participant in the qualitative phase related his age to the interpretation of the item "I've been feeling useful". Interaction terms between age and this item were therefore added to investigate the potential relationship between these two variables. Even though the relationships between other individual-specific covariates (i.e. participants' gender, ethnicity and education level) and C-TTO values were not documented in the results of the qualitative phase, these individual-specific covariates were added as dummy variables to explore potential group-based effects.

6.2.2. Conditional Logit model for the DCE data

The Conditional Logit model proposed by McFadden was used to model the data (McFadden, 1973). The 28 dummy variables included in this regression were the same as that of the Heteroskedastic Tobit model for the C-TTO. The dependent variable was a binary stated choice, with a value of 1 indicating the chosen option for each DCE pair.

Similar to the C-TTO model specification, the effect of incorporating different covariates was explored in addition to the main effects variables. However, it was noted that the incorporation of an individual-specific covariate in the conditional logit model was meaningless, as this would fall out of the probability of the chosen option (Greene, 2003). As a result, the covariates SWEMWBS score, gender, ethnicity, and education level were not considered. Deterministic heterogeneity was only explored by the interaction terms between age and the item “I’ve been feeling useful”, as informed by the result of the qualitative phase. Considering the limited modelling power due to small sample size, the interaction parameters between the item “I’ve been feeling useful” and other demographic variables were not explored. As none of the DCE responses was deleted in addition to the four missing answers, there were 2246 valid responses in total.

6.2.2.1. Rescaling

The utility values estimated from this model cannot be used in economic evaluation as they were estimated on a latent scale, but not estimated on a 0 to 1 scale for the use of value set generation. There are several main ways documented in the literature to rescale DCE values to a C-TTO comparable scale: anchoring using the coefficient for “dead”, anchoring the worst state using TTO, mapping DCE values onto TTO values, and constructing a hybrid model (Bahrampour *et al.*, 2020; Rowen *et al.*, 2015). The method of anchoring using the coefficient for “dead” was not possible in this study as the state “dead” was not included as an attribute in the DCE profile. The dead dummy variable was undefined in this case. Considering this, the other three methods were focused on in this chapter to rescale the DCE values.

6.2.2.1.1. *Anchoring to the lowest mental well-being state of the C-TTO*

The DCE latent scale was rescaled to the C-TTO comparable scale through anchoring the DCE value of the lowest mental well-being state (i.e. 1111111) at the C-TTO value of the lowest mental well-being state (i.e. 1111111). The underlying assumption of this method was that the DCE unscaled value is linearly proportional to the DCE rescaled value by a factor of α . Mathematically, the rescaling relationship can be generalised to the following formula:

$$DCE \text{ rescaled value for a state } i = \alpha DCE \text{ unscaled value for a state } i + \gamma \quad \dots (5)$$

$$\alpha = \frac{DCE \text{ rescaled value for a state } i - \gamma}{DCE \text{ unscaled value for a state } i}, \text{ where}$$

γ = a constant term. To anchor the DCE value of the lowest mental well-being state at the C-TTO value of the lowest mental well-being state, the *DCE rescaled value* was substituted by the C-TTO value generated by the Heteroskedastic Tobit model for the lowest mental well-being state. The *DCE unscaled value* was substituted by the DCE latent value generated by the Conditional Logit model for the lowest mental well-being state. It was noted that, by definition, the C-TTO value of the highest mental well-being was 1. The unscaled DCE value of the highest mental well-being state was normalised at 0. In this sense, the constant term γ was set at 1, with a view to ensuring that the resulting values generated by the rescaled DCE model and the C-TTO model for both the highest and lowest mental well-being states were the same. After calculating the α , the DCE rescaled coefficients for all attribute levels could then be obtained.

6.2.2.1.2. Mapping DCE onto C-TTO

The idea of this method is to derive the utility values for all C-TTO states based on the latent DCE values, using the following mapping function:

$$TTO = f(DCE) + \varepsilon \quad \dots (6), \text{ where}$$

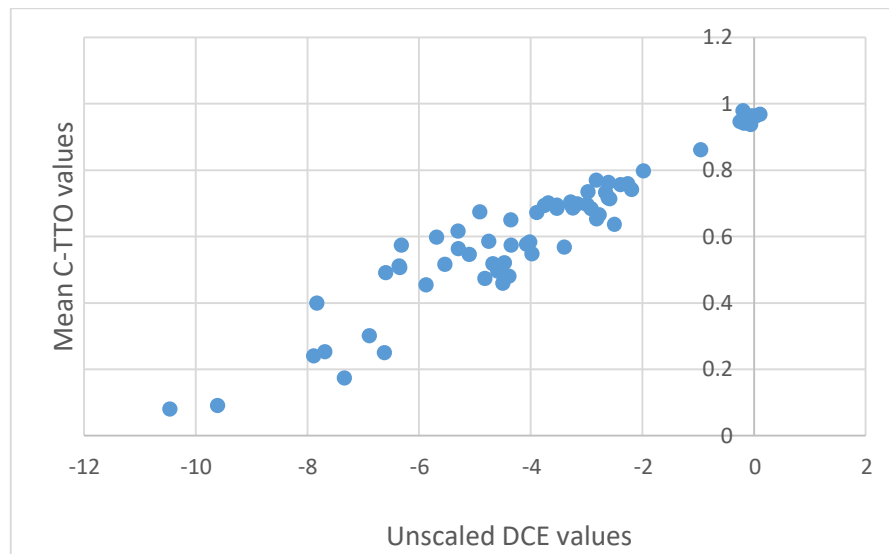
TTO = The mean C-TTO value of each of the 64 SWEMWBS states valued by the participants during the completion of C-TTO tasks.

DCE = The latent DCE utility values of each of these 64 SWEMWBS states.

ε = An error term of the model

It was sensible to assume an overall linear association between C-TTO values and latent DCE values, based on the following plot (Figure 11) between the mean C-TTO values and the latent DCE values for the 64 states. The OLS regression technique was used to model the rescaled factor.

Figure 11: Relationship between the mean C-TTO values and the unscaled DCE values for the 64 states valued by the participants



6.2.2.1.3. Hybrid model (the EuroQol hybrid model)

The hybrid model involves joint estimation of the C-TTO and DCE regression models. The underlying assumption is that the utility values derived from the C-TTO and the DCE responses may be different due to a unique utility function possessed by the respondents (Ramos-Goni *et al.*, 2017b). The coefficients for each of the attribute levels estimated from this hybrid model were based on maximum likelihood. Specifically, the likelihood function of a normal distribution for the C-TTO data was multiplied by the likelihood function of a conditional logit distribution for the DCE data (Oppe & Van Hout, 2010; Ramos-Goñi *et al.*, 2016; Ramos-Goni *et al.*, 2017b). As the DCE and C-TTO coefficients lay on a different scale, a rescaled parameter θ would be included for the optimisation of the joint distribution, with a view to allow the C-TTO and DCE models to differ by a monotonic transformation (i.e. TTO coefficient = θ *rescaled DCE coefficient).

The “hyreg” command in Stata developed by the EuroQol Group (Ramos-Goñi *et al.*, 2016) was used to derive the hybrid coefficients and the rescaling parameter θ . This hybrid method is named as the “EuroQol hybrid model” throughout this thesis.

6.2.3. Inverse Variance Weighting (IVW) approach for both the C-TTO and DCE data

This alternative hybrid approach (named the “IVW hybrid model” throughout this thesis) was adopted to derive weighted-average coefficients for the attribute levels from the modelled C-TTO and rescaled DCE coefficients (Lee *et al.*, 2016). The pooled coefficient

of an attribute level was calculated based on the assumption that the C-TTO and rescaled DCE coefficients were independent and normally distributed. In other words, the pooled coefficients were the weighted summation of the C-TTO and rescaled DCE coefficients. the variance of the pooled coefficient was the weighted summation of the C-TTO and rescaled DCE variances.

The formula for calculating the pooled coefficients and the pooled standard errors by this approach is as follows:

$$Pooled\ coefficient_{ij} = TTO_{ij} * \left[\frac{\frac{1}{(S_{TTO_{ij}})^2}}{\frac{1}{(S_{TTO_{ij}})^2} + \frac{1}{(S_{DCE_{ij}})^2}} \right] + DCE_{ij} * \left[\frac{\frac{1}{(S_{DCE_{ij}})^2}}{\frac{1}{(S_{TTO_{ij}})^2} + \frac{1}{(S_{DCE_{ij}})^2}} \right] \dots (7)$$

*Pooled Standard error*_{ij} =

$$\sqrt{\left[S_{TTO_{ij}} * \frac{\frac{1}{(S_{TTO_{ij}})^2}}{\frac{1}{(S_{TTO_{ij}})^2} + \frac{1}{(S_{DCE_{ij}})^2}} \right]^2 + \left[S_{DCE_{ij}} * \frac{\frac{1}{(S_{DCE_{ij}})^2}}{\frac{1}{(S_{TTO_{ij}})^2} + \frac{1}{(S_{DCE_{ij}})^2}} \right]^2} \dots (8)$$

, where

*Pooled coefficient*_{ij} = The weighted-average pooled coefficient for the attribute *i* with level *j*

*Pooled Standard error*_{ij} = The weighted-average pooled standard error for the attribute *i* with level *j*

*TTO*_{ij} = The coefficient for the attribute *i* with level *j* derived from the C-TTO model

*DCE*_{ij} = The rescaled coefficient for the attribute *i* with level *j* derived from the DCE model

*S*_{TTO_{ij}} = The standard error for the attribute *i* with level *j* derived from the C-TTO model

*S*_{DCE_{ij}} = The rescaled standard error for the attribute *i* with level *j* derived from the DCE model

6.2.4. Description of explanatory variables

Table 24 below provides a description of all explanatory variables included in the modelling of C-TTO and/or DCE data.

Table 24: A description of the explanatory variables

| Name of variable | Meaning | Reference level |
|------------------------------|--|---|
| <i>Main effect variables</i> | | |
| optimistic1 | None of the time feeling optimistic about the future | All of the time feeling optimistic about the future |
| optimistic2 | Rarely feeling optimistic about the future | |
| optimistic3 | Some of the time feeling optimistic about the future | |
| optimistic4 | Often feeling optimistic about the future | |
| useful1 | None of the time feeling useful | All of the time feeling useful |
| useful2 | Rarely feeling useful | |
| useful3 | Some of the time feeling useful | |
| useful4 | Often feeling useful | |
| relaxed1 | None of the time feeling relaxed | All of the time feeling relaxed |
| relaxed2 | Rarely feeling relaxed | |
| relaxed3 | Some of the time feeling relaxed | |
| relaxed4 | Often feeling relaxed | |
| dealingproblems1 | None of the time dealing with problems well | All of the time dealing with problems well |
| dealingproblems2 | Rarely dealing with problems well | |
| dealingproblems3 | Some of the time dealing with problems well | |
| dealingproblems4 | Often dealing with problems well | |
| thinkingclearly1 | None of the time thinking clearly | All of the time thinking clearly |
| thinkingclearly2 | Rarely thinking clearly | |
| thinkingclearly3 | Some of the time thinking clearly | |
| thinkingclearly4 | Often thinking clearly | |
| closetopeople1 | None of the time feeling close to other people | All of the time feeling close to other people |
| closetopeople2 | Rarely feeling close to other people | |

| | | |
|--------------------------|--|---|
| closetopeople3 | Some of the time feeling close to other people | |
| closetopeople4 | Often feeling close to other people | |
| makeupownmind1 | None of the time able to make up my own mind about things | All of the time able to make up my own mind about things |
| makeupownmind2 | Rarely able to make up my own mind about things | |
| makeupownmind3 | Some of the time able to make up my own mind about things | |
| makeupownmind4 | Often able to make up my own mind about things | |
| <i>Covariates</i> | | |
| SWEMWBS | Total score for the Short Warwick-Edinburgh Mental Wellbeing Scale | Not applicable |
| male | Male participants | Female participants |
| gender_others | Participants who declared their gender as “neutral” or preferred not to declare their gender. | |
| asian | Asian / Asian British | White |
| black | Black / African / Caribbean / Black British | |
| mixed | Mixed / Multiple ethnic groups | |
| ethnicity_others | Other ethnic group | |
| belowundergrad | Participants whose education level was below undergraduate: no education background, grammar school, GCSE, O-Level, A-Level, and diploma. | Participants whose education level was undergraduate or above: undergraduate, postgraduate (i.e. Master, PhD, Others) |
| education_others | Participants who did not provide sufficient information about their highest education level (i.e. either below undergraduate or undergraduate/above), e.g. some only declared their highest education level as “professional qualification”. | |
| age | The age of participants | Not applicable |

| | | |
|-------------|--|---|
| useful1*age | An interaction term between “none of the time feeling useful” and the age of participants. | An interaction term between “all of the time feeling useful” and the age of participants. |
| useful2*age | An interaction term between “rarely feeling useful” and the age of participants. | |
| useful3*age | An interaction term between “some of the time feeling useful” and the age of participants. | |
| useful4*age | An interaction term between “often feeling useful” and the age of participants. | |

6.2.5. Model analysis

The criteria for the assessment of model performance were as follows:

- Logical consistency of the estimators: A lower (higher) level of an item should theoretically generate a higher (lower) utility decrement, *ceteris paribus*.
- The statistical significance of the estimators.
- Goodness of fit: This was assessed by the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) to examine the efficiency of models, with an attempt to strike balance between the sophistication (complexity) and the simplicity of the model. Also, a Wald test was conducted to explore the addition of covariates on the model fit improvement.
- Parsimony: A sophisticatedly simple model with optimal number of predictors (without overfitting) was sufficient to explain the model well.
- Mean absolute deviation (MAD) and Root Mean Square Deviation (RMSD): As the idea of the two DCE rescaling methods was to convert the DCE values from a latent scale to a C-TTO comparable scale, the MAD and RMSD of these two methods calculated using the following formulae were compared to investigate their ability to predict the observed C-TTO values for the 64 states valued by the participants (Rowen *et al.*, 2015).

$$MAD = \frac{\sum_{i=1}^{64} |x_i - \mu_i|}{64} \quad \dots (9), \text{ where}$$

x_i = The predicted DCE rescaled utility value based on anchoring/mapping for state i .

μ_i = The mean observed C-TTO value for state i.

The number of predictions/states with rescaled utility values greater than 0.05, 0.1 and 0.2 from observed mean C-TTO value was also investigated.

$$RMSD = \sqrt{\frac{\sum_{j=1}^{2103} (x_j - y_j)^2}{2103}} \dots (10), \text{ where}$$

x_j = The observed C-TTO value for the observation j, given that there were 2103 observations in total for the 64 states valued by the participants.

y_j = The predicted DCE rescaled utility value based on anchoring/mapping for observation j.

Additionally, to discover the relative importance of the seven attributes to the participants, the ranking of the attributes was ordered, based on comparing the sum of the absolute value of the level coefficients for each attribute.

6.2.6. Sensitivity analysis

The robustness of the C-TTO and DCE models was further explored by sensitivity analyses. For the modelling of the C-TTO data, the estimated parameters were investigated when the values of the flagged C-TTO states in the Feedback Module were included. For the DCE data, the 40 responses of the four participants with potential strategic/repeated pattern or sign of lacking engagement (quanti034: BBBAABBBAA, quanti099: BBBBAAAAA, quanti191: AAAAAAAAAAA, quanti193: ABAABABAAB) were excluded.

6.3. Results

The C-TTO and DCE main effects models were separately produced, followed by adding covariates and interaction terms to the main effects models.

6.3.1. The C-TTO models

Table 25: Results from different specifications of the C-TTO data

| Model 1A: | Model 1B: | Model 1C: |
|------------------|----------------------|----------------------|
| Main effects | Main effects + | Main effects + |
| | SWEMWBS score + | SWEMWBS score + |
| | Gender + Ethnicity + | Gender + Ethnicity + |
| | Education | Education + Age + |

| | Interaction terms | | | | | |
|------------------|-----------------------------|-------|----------------------------|-------|---------------------------|-------|
| | b (R.SE) | p | b (R.SE) | p | b (R.SE) | p |
| optimistic1 | -0.153*** (0.036) | 0.000 | -0.148*** (0.031) | 0.000 | -0.134*** (0.030) | 0.000 |
| optimistic2 | -0.135*** (0.034) | 0.000 | -0.117*** (0.028) | 0.000 | -0.107*** (0.028) | 0.000 |
| optimistic3 | -0.058 (0.036) | 0.102 | -0.050* (0.030) | 0.094 | -0.047 (0.029) | 0.110 |
| optimistic4 | -0.072** (0.031) | 0.019 | -0.058** (0.028) | 0.038 | -0.051* (0.027) | 0.061 |
| useful1 | -0.144*** (0.026) | 0.000 | -0.131*** (0.023) | 0.000 | -0.269*** (0.072) | 0.000 |
| useful2 | -0.117*** (0.034) | 0.001 | -0.082*** (0.029) | 0.004 | -0.045 (0.079) | 0.572 |
| useful3 | -0.032 (0.035) | 0.369 | -0.032 (0.029) | 0.272 | -0.006 (0.074) | 0.937 |
| useful4 | -0.036 (0.030) | 0.231 | -0.020 (0.026) | 0.455 | -0.067 (0.072) | 0.354 |
| relaxed1 | -0.168*** (0.035) | 0.000 | -0.158*** (0.029) | 0.000 | -0.155*** (0.029) | 0.000 |
| relaxed2 | -0.189*** (0.035) | 0.000 | -0.144*** (0.032) | 0.000 | -0.144*** (0.031) | 0.000 |
| relaxed3 | -0.089** (0.038) | 0.019 | -0.077** (0.034) | 0.025 | -0.061* (0.033) | 0.067 |
| relaxed4 | -0.066* (0.034) | 0.053 | -0.052* (0.029) | 0.077 | -0.047 (0.029) | 0.103 |
| dealingproblems1 | -0.097*** (0.036) | 0.007 | -0.120*** (0.034) | 0.000 | -0.121*** (0.033) | 0.000 |
| dealingproblems2 | -0.106*** (0.033) | 0.001 | -0.100*** (0.028) | 0.000 | -0.096*** (0.029) | 0.001 |
| dealingproblems3 | -0.069** (0.035) | 0.047 | -0.080*** (0.029) | 0.007 | -0.078*** (0.029) | 0.006 |
| dealingproblems4 | -0.018 (0.027) | 0.490 | -0.031 (0.025) | 0.211 | -0.021 (0.025) | 0.402 |
| thinkingclearly1 | -0.204*** | 0.000 | -0.174*** | 0.000 | -0.171*** | 0.000 |

| | | | | | | |
|------------------|------------------|-------|-----------|-------|-----------|-------|
| | (0.038) | | (0.035) | | (0.035) | |
| thinkingclearly2 | -0.107*** | 0.001 | -0.077*** | 0.004 | -0.076*** | 0.005 |
| | (0.032) | | (0.027) | | (0.027) | |
| thinkingclearly3 | -0.046 | 0.229 | -0.049 | 0.169 | -0.063* | 0.073 |
| | (0.038) | | (0.036) | | (0.035) | |
| thinkingclearly4 | -0.016 | 0.601 | -0.019 | 0.476 | -0.023 | 0.394 |
| | (0.030) | | (0.026) | | (0.027) | |
| closetopeople1 | -0.169*** | 0.000 | -0.193*** | 0.000 | -0.185*** | 0.000 |
| | (0.031) | | (0.028) | | (0.028) | |
| closetopeople2 | -0.169*** | 0.000 | -0.142*** | 0.000 | -0.140*** | 0.000 |
| | (0.036) | | (0.031) | | (0.031) | |
| closetopeople3 | -0.073** | 0.026 | -0.089*** | 0.002 | -0.082*** | 0.004 |
| | (0.033) | | (0.029) | | (0.028) | |
| closetopeople4 | 0.000 | 0.994 | -0.021 | 0.413 | -0.022 | 0.396 |
| | (0.028) | | (0.026) | | (0.026) | |
| makeupownmind1 | -0.165*** | 0.000 | -0.150*** | 0.000 | -0.147*** | 0.000 |
| | (0.034) | | (0.029) | | (0.029) | |
| makeupownmind2 | -0.121*** | 0.000 | -0.101*** | 0.000 | -0.094*** | 0.000 |
| | (0.031) | | (0.027) | | (0.026) | |
| makeupownmind3 | -0.082** | 0.024 | -0.057* | 0.066 | -0.059* | 0.053 |
| | (0.036) | | (0.031) | | (0.030) | |
| makeupownmind4 | -0.021 | 0.494 | -0.038 | 0.176 | -0.015 | 0.603 |
| | (0.030) | | (0.028) | | (0.029) | |
| SWEMWBS | | | 0.009*** | 0.000 | 0.010*** | 0.000 |
| | | | (0.002) | | (0.002) | |
| male | | | 0.001 | 0.927 | 0.005 | 0.732 |
| | | | (0.016) | | (0.016) | |
| gender_others | | | -0.292*** | 0.001 | -0.270*** | 0.004 |
| | | | (0.090) | | (0.093) | |
| asian | | | -0.062** | 0.012 | -0.089*** | 0.001 |
| | | | (0.025) | | (0.028) | |
| black | | | -0.127*** | 0.002 | -0.136*** | 0.000 |
| | | | (0.041) | | (0.038) | |
| mixed | | | -0.088 | 0.170 | -0.086 | 0.182 |
| | | | (0.064) | | (0.064) | |
| ethnicity_others | | | 0.164*** | 0.000 | 0.144*** | 0.001 |

| | | | | | | |
|------------------|----------|-------|-----------|-------|-----------|-------|
| | | | (0.035) | | (0.042) | |
| belowundergrad | | | -0.106*** | 0.000 | -0.104*** | 0.000 |
| | | | (0.022) | | (0.022) | |
| education_others | | | -0.005 | 0.941 | 0.009 | 0.892 |
| | | | (0.063) | | (0.065) | |
| age | | | | | -0.001 | 0.157 |
| | | | | | (0.001) | |
| useful1*age | | | | | 0.003** | 0.050 |
| | | | | | (0.001) | |
| useful2*age | | | | | -0.001 | 0.653 |
| | | | | | (0.002) | |
| useful3*age | | | | | -0.001 | 0.627 |
| | | | | | (0.001) | |
| useful4*age | | | | | 0.001 | 0.485 |
| | | | | | (0.001) | |
| constant | 1.174*** | 0.000 | 0.950*** | 0.000 | 0.972*** | 0.000 |
| | (0.033) | | (0.052) | | (0.068) | |
| AIC | 2326.403 | | 2041.653 | | 2001.570 | |
| BIC | 2654.168 | | 2471.139 | | 2487.566 | |
| N | 2103.000 | | 2103.000 | | 2103.000 | |

Number of statistically significant main effects parameters at 5% level 19

18

15

Number of statistically significant main effects parameters at 10% level 20

21

19

Notes: *** significant at 1%, **significant at 5%, * significant at 10%

Robust standard errors are presented in parentheses.

Potentially logical inconsistent coefficients are highlighted in bold.

AIC indicates Akaike Information Criterion; BIC, Bayesian Information Criterion; N, number of observations

In Table 25, the main effects model is labelled as Model 1A. At first glance, the sign of coefficients was at the expected negative direction, indicating mental health disutility from the reference level “all of the time”. The number of statistically insignificant main effects parameters at the 5% level was nine, one of which one was statistically significant at the 10% level. The problem of potentially logical inconsistency was discovered for eleven coefficients (highlighted in bold). For example, the utility decrement for “often feeling optimistic about the future” (-0.072) was higher than the utility decrement for “some of the time feeling optimistic about the future” (-0.058). It was also interesting to realise that the coefficient for the variable “closetopeople4” was statistically insignificant at zero. This implied the possibility of a higher utility for this mental well-being level over the highest possible level of mental well-being.

In addition to the main effects model, the covariates containing the SWEMWBS score, gender, ethnicity, and education backgrounds for each participant were added in Model 1B. The covariate “SWEMWBS” was positive and statistically significant, revealing that one unit increase (decrease) in SWEMWBS score was associated with a 0.009 increase (decrease) in the predicted C-TTO value. For the dummy variables of gender covariate, compared to the female participants, the effect of male participants on the predicted C-TTO value was statistically insignificant. However, the variable “gender_others” was negative (-0.292) and statistically significant at the 1% level, even though the result was derived by the valuation responses from only three participants in this demographic group. This indicated that the three participants who declared their gender as neutral or did not disclose their gender generally gave lower predicted C-TTO value, compared to the female participants. For the covariates related to ethnicity, compared to White respondents, Asian and Black participants reported lower predicted C-TTO values. These were indicated by the negative and statistically significant coefficients for the variables “asian” (-0.062) and “black” (-0.127). Also, the two participants who declared themselves as “Other ethnic group” were associated with higher predicted C-TTO value, compared to “White”. No statistically significant effect on predicted C-TTO value was observed for the participants with mixed ethnicities. Moreover, the covariate related to the participants’ education background was investigated. Compared to the participants with higher education level (i.e. undergraduate and postgraduate), participants with lower education level (i.e. below undergraduate) were associated with lower predicted C-TTO value. This relationship was indicated by the negative and statistically significant coefficient for the variable

“belowundergrad” (-0.106) at the 1% level. The effect of participants with “education_others” on the predicted C-TTO value was statistically insignificant.

In terms of the performance of the main effects parameters in Model 1B, the signs for coefficients were all negative. Compared to Model 1A, 10 out of the 28 main effects coefficients were statistically insignificant at the 5% level, whilst three coefficients were significant at the 10% level. The number of potentially logical inconsistent coefficients decreased from 11 to 2 (optimistic3 and optimistic4). The explanatory power of this model was also better than Model 1A, as indicated by a 12.24% decrease in AIC and a 6.896% decrease in BIC. The improvement in goodness of fit was also supported by the result of the Wald test, as shown in Table 26. The p-value associated with the chi-squared test indicated that the null hypothesis of zero coefficient for the SWEMWBS, gender, ethnicity, and education covariates was rejected. This means that the inclusion of these variables resulted in a statistically significant improvement in the model fit. All these diagnostics are suggestive of superiority of Model 1B over Model 1A.

Over and above this, the covariates of participants’ age and the interaction terms between age and the item “I’ve been feeling useful” were added to Model 1B to create Model 1C. The result showed that the coefficients for the covariates related to the SWEMWBS score, gender, ethnicity, and education level performed similarly as Model 1B. Consistent with Model 1B, the coefficients “SWEMWBS” and “ethnicity_others” were positive and statistically significant at 1% level in Model 1C. The coefficients “gender_others”, “asian”, “black”, and “belowundergrad” were negative and statistically significant at 1% level. Particularly, the significant level for “asian” changed from 5% in Model 1B to 1% in Model 1C. The magnitude change (-0.027) for the coefficient “asian” was also the largest among all statistically significant covariates, with a decrease in coefficient from -0.062 in Model 1B to -0.089 in Model 1C. The covariates “Age”, “useful2*Age”, “useful3*Age” and “useful4*Age” were statistically insignificant, revealing the absence of association between the C-TTO values and these variables. However, the interaction term “useful1*Age” was positive and statistically significant at the 5% level. This indicates that the higher the age, the greater the effect of “none of the time feeling useful” on the C-TTO values. However, the inclusion of these interaction terms increased the number of statistically insignificant main effect coefficients at the 5% from 10 in Model 1B to 13 in Model 1C. The number of statistically significant main effect coefficients at the 10% level

also dropped from 21 in Model 1B to 19 in Model 1C. Moreover, the number of potentially logical inconsistent main effect coefficients increased from 2 in Model 1B to 3 in Model 1C. Furthermore, the goodness-of-fit performance evaluated by the AIC and BIC statistics was ambiguous, as the changes in these two statistics moved in different directions. Compared to Model 1B, the AIC of Model 1C decreased by 1.963%, whereas the BIC increased by 0.665%. However, it should be noted that the result of the Wald test in Table 26 showed a statistically significant improvement in the model fit of this model over Model 1B, meaning that the inclusion of these additional variables in Model 1C played a role in explaining the outputs of the model. Nevertheless, Model 1B was preferred to Model 1C based on the principle of parsimony. Even though the result of the Wald test supported the improvement in model fit for Model 1C, the percentage changes of the AIC and BIC statistics were negligible. The increase in BIC statistic even suggested a decrease in goodness of fit. Also, there were more logically inconsistent and statistically insignificant coefficients in Model 1C, worsening its overall performance. Considering these results, Model 1B was identified as possessing the optimal number of explanatory variables that explain the model well.

Based on the above analysis, given that Model 1B was preferred to Model 1A and Model 1C, Model 1B was regarded as the most preferred model overall for the C-TTO data. However, in terms of value set calculation, Model 1A was used instead of Model 1B to address the limitations of the DCE model specification and the rescaling of the DCE coefficients based on anchoring. These limitations will be followed up in sections 6.3.2.1 and 6.3.5.

The summation of the absolute values of the level coefficients for each attribute in Model 1A is reported in Table 27. As the magnitude of a coefficient represents the weight attached to an attribute level, summing the level coefficients for an attribute is a close proxy for indicating the level of importance for this attribute. The results suggest that the ranking of the attributes based on this summation method was ordered as: relaxed (the most important attribute), optimistic, closetopeople, makeupownmind, thinkingclearly, useful, and dealingproblems (the least important attribute). Among the coefficients of all attribute levels, thinkingclearly1 (I've been none of the time thinking clearly) received the highest weight in absolute value (0.204), indicating the largest change from the base level 5. The attribute level with the least weight in absolute value was closetopeople4 (I've been

often feeling close to other people) (0.000227), indicating the smallest change from base level 5.

Table 26: Wald tests for model comparison

| Model comparison | chi-squared | p-value |
|------------------------|-------------|----------|
| Model 1A v.s. Model 1B | 252.23 | 0.000*** |
| Model 1B v.s. Model 1C | 23.02 | 0.011** |
| Model 2A v.s. Model 2B | 3.77 | 0.439 |

Notes: *** significant at 1%, **significant at 5%.

Table 27: The total weight for each attribute in the Model 1A

| Rank | Attribute | Total weight in absolute value |
|------|--|--|
| 1 | I've been feeling relaxed | $0.168 + 0.189 + 0.089 + 0.066 = 0.512$ |
| 2 | I've been feeling optimistic about the future | $0.153 + 0.135 + 0.058 + 0.072 = 0.418$ |
| 3 | I've been feeling close to other people | $0.169 + 0.169 + 0.073 + 0.000227 = 0.411$ |
| 4 | I've been able to make up my own mind about things | $0.165 + 0.121 + 0.082 + 0.021 = 0.389$ |
| 5 | I've been thinking clearly | $0.204 + 0.107 + 0.046 + 0.016 = 0.373$ |
| 6 | I've been feeling useful | $0.144 + 0.117 + 0.032 + 0.036 = 0.329$ |
| 7 | I've been dealing with problems well | $0.097 + 0.106 + 0.069 + 0.018 = 0.29$ |

6.3.2. The DCE models

Table 28: Results from different specifications of the DCE data

| | Model 2A: | | | Model 2B: | |
|-------------|----------------------|-------|---|----------------------------------|-------|
| | Main effects | | <i>rescaled b</i> (<i>anchoring</i>) | Main effects + interaction terms | |
| | b (R.SE) | p | | b (R.SE) | p |
| optimistic1 | -1.502*** (0.274) | 0.000 | -0.158*** (0.029) | -1.492*** (0.275) | 0.000 |
| optimistic2 | -1.028*** (0.224) | 0.000 | -0.108*** (0.024) | -1.010*** (0.225) | 0.000 |
| optimistic3 | -0.267 | 0.114 | -0.028 | -0.261 | 0.124 |

| | | | | | |
|------------------|--------------|-------|--------------|--------------|-------|
| | (0.169) | | (0.018) | (0.170) | |
| optimistic4 | -0.061 | 0.532 | -0.006 | -0.064 | 0.512 |
| | (0.098) | | (0.01) | (0.098) | |
| useful1 | -1.443*** | 0.000 | -0.152*** | -1.738*** | 0.000 |
| | (0.265) | | (0.028) | (0.435) | |
| useful2 | -1.071*** | 0.000 | -0.113*** | -0.805** | 0.047 |
| | (0.207) | | (0.022) | (0.404) | |
| useful3 | -0.435*** | 0.006 | -0.046*** | -0.492 | 0.190 |
| | (0.158) | | (0.017) | (0.375) | |
| useful4 | -0.177* | 0.051 | -0.019* | -0.208 | 0.509 |
| | (0.091) | | (0.01) | (0.315) | |
| relaxed1 | -1.369*** | 0.000 | -0.144*** | -1.350*** | 0.000 |
| | (0.262) | | (0.028) | (0.263) | |
| relaxed2 | -0.819*** | 0.000 | -0.086*** | -0.812*** | 0.000 |
| | (0.198) | | (0.021) | (0.198) | |
| relaxed3 | -0.311** | 0.040 | -0.033** | -0.298** | 0.049 |
| | (0.151) | | (0.016) | (0.151) | |
| relaxed4 | 0.028 | 0.762 | 0.003 | 0.023 | 0.802 |
| | (0.093) | | (0.01) | (0.093) | |
| dealingproblems1 | -1.392*** | 0.000 | -0.146*** | -1.378*** | 0.000 |
| | (0.278) | | (0.029) | (0.279) | |
| dealingproblems2 | -0.960*** | 0.000 | -0.101*** | -0.950*** | 0.000 |
| | (0.208) | | (0.022) | (0.208) | |
| dealingproblems3 | -0.474*** | 0.004 | -0.05*** | -0.460*** | 0.005 |
| | (0.164) | | (0.017) | (0.165) | |
| dealingproblems4 | -0.196** | 0.038 | -0.021** | -0.203** | 0.032 |
| | (0.094) | | (0.01) | (0.095) | |
| thinkingclearly1 | -1.261*** | 0.000 | -0.133*** | -1.244*** | 0.000 |
| | (0.249) | | (0.026) | (0.250) | |
| thinkingclearly2 | -0.843*** | 0.000 | -0.089*** | -0.833*** | 0.000 |
| | (0.193) | | (0.02) | (0.193) | |
| thinkingclearly3 | -0.168 | 0.239 | -0.018 | -0.163 | 0.256 |
| | (0.143) | | (0.015) | (0.144) | |
| thinkingclearly4 | 0.105 | 0.286 | 0.011 | 0.105 | 0.287 |
| | (0.098) | | (0.01) | (0.099) | |
| closetopeople1 | -2.239*** | 0.000 | -0.235*** | -2.228*** | 0.000 |

| | | | | | |
|--|-----------|-------|-----------|-----------|-------|
| | (0.273) | | (0.029) | (0.274) | |
| closetopeople2 | -1.539*** | 0.000 | -0.162*** | -1.524*** | 0.000 |
| | (0.209) | | (0.022) | (0.210) | |
| closetopeople3 | -0.635*** | 0.000 | -0.067*** | -0.628*** | 0.000 |
| | (0.149) | | (0.016) | (0.150) | |
| closetopeople4 | -0.247** | 0.010 | -0.026** | -0.243** | 0.012 |
| | (0.096) | | (0.01) | (0.097) | |
| makeupownmind1 | -1.254*** | 0.000 | -0.132*** | -1.244*** | 0.000 |
| | (0.270) | | (0.028) | (0.269) | |
| makeupownmind2 | -0.596*** | 0.003 | -0.063*** | -0.581*** | 0.004 |
| | (0.201) | | (0.021) | (0.201) | |
| makeupownmind3 | -0.469*** | 0.002 | -0.049*** | -0.465*** | 0.002 |
| | (0.152) | | (0.016) | (0.152) | |
| makeupownmind4 | -0.036 | 0.694 | -0.004 | -0.041 | 0.650 |
| | (0.091) | | (0.01) | (0.091) | |
| useful1*Age | | | | 0.006 | 0.389 |
| | | | | (0.007) | |
| useful2*Age | | | | -0.005 | 0.448 |
| | | | | (0.007) | |
| useful3*Age | | | | 0.001 | 0.848 |
| | | | | (0.007) | |
| useful4*Age | | | | 0.001 | 0.916 |
| | | | | (0.006) | |
| constant_option A | -0.095** | 0.039 | -0.01** | -0.098** | 0.033 |
| | (0.046) | | (0.005) | (0.046) | |
| AIC | 2849.381 | | | 2853.634 | |
| BIC | 3035.273 | | | 3065.166 | |
| N | 2246 | | | 2246 | |
| Number of statistically significant main effects parameters at 5% level | 21 | | | 20 | |
| Number of statistically significant main effects parameters at 10% level | 22 | | | 20 | |

Notes: *** significant at 1%, **significant at 5%, * significant at 10%

Robust standard errors are presented in parentheses.

Potentially logical inconsistent coefficients are highlighted in bold.

AIC indicates Akaike Information Criterion; BIC, Bayesian Information Criterion; N, number of observations.

The main effects model is labelled as Model 2A in Table 28. In general, the sign of the marginal utility for all main effects coefficients was negative, except for the coefficients of “relaxed4” and “thinkingclearly4” (highlighted in bold). Although the positive coefficients of these two variables were statistically insignificant, these could indicate a potential increase in utility when changing from feeling relaxed all of the time to often feeling relaxed, and from thinking clearly all of the time to often thinking clearly. Other than these two coefficients, there were no other logically inconsistent coefficients identified in this model. The number of statistically insignificant coefficients at the 5% level was seven, of which one of them (useful4) became significant at the 10% level. The coefficient for the constant term was negative and statistically significant at the 5% level. This implied that there was a left-right bias, in the sense that participants tended to choose the alternative displayed on the right-hand side (i.e. option B).

In Model 2B, interaction terms between the participants’ age and the attribute “I’ve been feeling useful” were added in addition to the main effects parameters. Identical to Model 2A, there were two logically inconsistent and statistically insignificant coefficients (highlighted in bold). The number of statistically insignificant coefficients at the 5% level increased from seven in Model 2A to eight in this model. No extra statistically significant coefficient was identified at the 10% significance level. All interaction terms in this model were statistically insignificant, contributing no meaningful information in explaining the model. The increase in AIC and BIC statistics by 0.15% and 0.98% respectively also indicated a poor goodness of fit for this model. Moreover, the Wald test for model comparison between Model 2A and Model 2B was performed. The result shown in Table 26 was not statistically significant, meaning that the null hypothesis of simultaneously zero coefficients for the interaction terms was not rejected. The inclusion of these

interaction terms therefore did not improve the model fit. All these statistics consistently supported Model 2A as the most preferred model for the DCE data.

6.3.2.1. Anchoring to the lowest mental well-being state of the C-TTO

The rescaled coefficients in Model 2A derived from anchoring the DCE value of the lowest mental well-being state at the C-TTO value of the lowest mental well-being state are also reported in Table 28. As mentioned in Equation 5 of the methods section, the process of anchoring required the calculation of α by the following equation:

$$\alpha = \frac{C - TTO \text{ utility value for the state } 1111111 - 1}{DCE \text{ unscaled utility value for the state } 1111111}$$

Although Model 1B was preferred to the Model 1A, the model coefficients for the main effects variables were very similar. To ensure the compatibility of the model variables and given that the individual-specific SWEMWBS score covariate could not be explored within the DCE model, Model 1A for the C-TTO data alongside Model 2A for the DCE were subsequently used for value set calculation and anchoring of the DCE values.

The C-TTO utility value generated by Model 1A for state 1111111 = 1 – 0.153 – 0.144 – 0.168 – 0.097 – 0.204 – 0.169 – 0.165 = –0.1

The DCE unscaled utility value generated by Model 2A for state 1111111 = 0 – 1.502 – 1.443 – 1.369 – 1.392 – 1.261 – 2.239 – 1.254 = –10.46

$$\alpha = \frac{-0.1 - 1}{-10.46} = 0.105$$

Once the α is derived, it can then be applied into the generalised formula to rescale any DCE utility value for a particular state.

$$DCE \text{ rescaled value for a state} = \alpha DCE \text{ unscaled value for a state} + \gamma$$

For example, the DCE rescaled utility for state 1111111 should now be equal to the C-TTO utility value for state 1111111, i.e. the DCE rescaled utility value in Model 2A for state 1111111 = –10.46 * 0.105 + 1 = –0.1 = the C-TTO utility value in Model 1A for state 1111111.

To calculate the rescaled DCE coefficients for all the attribute levels in Table 28, the following formula was applied,

$$1 + \alpha \sum x_{ij} = \text{rescaled utility value of a particular state}$$

where x_{ij} represents the unscaled coefficient for attribute $i \in [1,7]$ at level $j \in [1,4]$. The term $\alpha \sum x_{ij}$ is simply the rescaled coefficient for attribute i at level j . For example, the rescaled coefficient for optimistic1 = $-1.502 * 0.105 = -0.158$.

The summation of the absolute value of the rescaled level coefficients for each attribute based on anchoring in Model 2A is reported in Table 29. The results suggest that the ranking of the attributes based on this summation method was ordered as: closetopeople (the most important attribute), useful, dealingproblems, optimistic, relaxed, thinkingclearly, and makeupownmind (the least important attribute). Among the coefficients of all attribute levels, closetopeople1 (I've been none of the time feeling close to other people) received the highest weight in absolute value (0.235). The attribute level with the least weight in absolute value was relaxed4 (I've been often feeling relaxed) (0.003).

Table 29: The total weight for each attribute in the Model 2A based on anchoring

| Rank | Attribute | Total weight in absolute value |
|------|--|---|
| 1 | I've been feeling close to other people | $0.235 + 0.162 + 0.067 + 0.026 = 0.49$ |
| 2 | I've been feeling useful | $0.152 + 0.113 + 0.046 + 0.019 = 0.329$ |
| 3 | I've been dealing with problems well | $0.146 + 0.101 + 0.05 + 0.021 = 0.318$ |
| 4 | I've been feeling optimistic about the future | $0.158 + 0.108 + 0.028 + 0.006 = 0.301$ |
| 5 | I've been feeling relaxed | $0.144 + 0.086 + 0.033 + 0.003 = 0.266$ |
| 6 | I've been thinking clearly | $0.133 + 0.089 + 0.018 + 0.011 = 0.25$ |
| 7 | I've been able to make up my own mind about things | $0.132 + 0.063 + 0.049 + 0.004 = 0.248$ |

Regarding the predictive power of this method, the MAD was 0.067. The number of predictions with absolute deviation greater than 0.05, 0.1 and 0.2 from the observed mean C-TTO values was 33, 13 and 2, respectively. The RMSD was 0.375.

6.3.2.2. Mapping DCE onto C-TTO

Table 30 shows the mapping result generated by regressing the mean C-TTO value on the DCE unscaled values for the 64 SWEMWBS states.

Table 30: Mapping result

| | b (R.SE) | p |
|----------------------|--------------------|-------|
| DCE unscaled utility | 0.084*** (0.00) | 0.000 |
| constant | 0.954*** (0.01) | 0.000 |
| N | 64 | |

Notes: *** significant at 1%

Robust standard errors are presented in parentheses.

The coefficients of unscaled DCE utility values and the constant term were statistically significant at the 1% level. These coefficients could then be applied to calculate the rescaled DCE values from the unscaled DCE values, based on the following formula:

$$\begin{aligned} & \text{DCE rescaled utility value for a state } i \\ & = \text{DCE unscaled utility value for a state } i * 0.084 + 0.954 \end{aligned}$$

For example, the utility value for the state 5555555 = 0*0.084 + 0.954 = 0.954.

The utility change of each attribute level from the highest level (i.e. level 5) could be calculated to explore the importance of each of the seven attributes. As an illustration, to calculate the weights of levels 1-4 relative to level 5 for the attribute makeupownmind, the rescaled utility value for the state 5555555 could be compared with the values of the states 5555551, 5555552, 5555553, 5555554. In other words, the difference between level 5 and each of the levels 1-4 could be derived by keeping the levels of the remaining six attributes constant. Table 31 demonstrates the calculation process of this example.

Table 31: Calculation of utility change for the levels of attribute makeupownmind

| State | Rescaled utility value | Utility decrement from level 5 |
|---------|------------------------|--------------------------------|
| 5555555 | 0.954 | 0 |
| 5555551 | 0.848664 | 0.848664 - 0.954 = -0.105 |

| | | |
|---------|----------|---------------------------|
| 5555552 | 0.903936 | 0.903936 - 0.954 = -0.05 |
| 5555553 | 0.914604 | 0.914604 - 0.954 = -0.039 |
| 5555554 | 0.950976 | 0.950976 - 0.954 = -0.003 |

This result showed that the utility decrements from state 5555555 for the attribute levels makeupownmind1, makeupownmind2, makeupownmind3, makeupownmind4 were - 0.105, -0.05, -0.039 and -0.003, respectively.

Through the application of this method, the weight of utility change from level 5 for all attribute levels can be derived and the results are presented in Table 32:

Table 32: Utility change from level 5 for all attribute levels

| Attribute level | Utility change from level 5 |
|------------------|-----------------------------|
| optimistic1 | -0.126 |
| optimistic2 | -0.086 |
| optimistic3 | -0.022 |
| optimistic4 | -0.005 |
| useful1 | -0.121 |
| useful2 | -0.09 |
| useful3 | -0.037 |
| useful4 | -0.015 |
| relaxed1 | -0.115 |
| relaxed2 | -0.069 |
| relaxed3 | -0.026 |
| relaxed4 | 0.002 |
| dealingproblems1 | -0.117 |
| dealingproblems2 | -0.081 |
| dealingproblems3 | -0.04 |
| dealingproblems4 | -0.016 |
| thinkingclearly1 | -0.106 |
| thinkingclearly2 | -0.071 |
| thinkingclearly3 | -0.014 |
| thinkingclearly4 | 0.009 |
| closetopeople1 | -0.188 |
| closetopeople2 | -0.129 |

| | |
|----------------|--------|
| closetopeople3 | -0.053 |
| closetopeople4 | -0.021 |
| makeupownmind1 | -0.105 |
| makeupownmind2 | -0.05 |
| makeupownmind3 | -0.039 |
| makeupownmind4 | -0.003 |

Note: Potentially logical inconsistent values are highlighted in bold.

The summation of the absolute value of the utility change from the best level for each attribute based on mapping in Model 2A is reported in Table 33. Identical to the weighting distribution of the attribute levels based on anchoring, the result suggested that the ranking of the attributes based on this summation method was ordered as: closetopeople (the most important attribute), useful, dealingproblems, optimistic, relaxed, thinkingclearly, and makeupownmind (the least important attribute). Among all attribute levels, closetopeople1 (I've been none of the time feeling close to other people) received the highest weight in absolute value (0.188). The attribute level with the least weight in absolute value was relaxed4 (I've been often feeling relaxed) (0.002).

Table 33: The total weight for each attribute in the Model 2A based on mapping

| Rank | Attribute | Total weight in absolute value |
|------|--|---|
| 1 | I've been feeling close to other people | $0.188 + 0.129 + 0.053 + 0.021 = 0.391$ |
| 2 | I've been feeling useful | $0.121 + 0.09 + 0.037 + 0.015 = 0.263$ |
| 3 | I've been dealing with problems well | $0.117 + 0.081 + 0.04 + 0.016 = 0.254$ |
| 4 | I've been feeling optimistic about the future | $0.126 + 0.086 + 0.022 + 0.005 = 0.24$ |
| 5 | I've been feeling relaxed | $0.115 + 0.069 + 0.026 + 0.002 = 0.212$ |
| 6 | I've been thinking clearly | $0.106 + 0.071 + 0.014 + 0.009 = 0.2$ |
| 7 | I've been able to make up my own mind about things | $0.105 + 0.05 + 0.039 + 0.003 = 0.198$ |

Lastly, the MAD for this method was 0.05. The number of predictions with absolute deviation greater than 0.05 and 0.1 from the observed mean C-TTO values was 29 and 10, respectively. The RMSD was 0.367.

6.3.2.3. The EuroQol hybrid model

The application of this method for deriving the hybrid coefficients and rescaled DCE coefficients for the SWEMWBS attribute levels is documented in Appendix 24. However, due to the failure to achieve an informative rescaling θ given the C-TTO and DCE responses, the hybrid coefficients for value set generation and the rescaling θ generated by this method were not adopted in this thesis. Instead, the result of an alternative hybrid approach for deriving the valuation set through combining the C-TTO and DCE data will be discussed in the following section.

6.3.3. The IVW hybrid model

Table 34 below shows the pooled coefficients and the corresponding standard errors for the attribute levels.

Table 34: Result of the IVW hybrid model

| Model 3: | |
|-----------------|---------------------|
| Main effects | |
| | b (R.SE) |
| optimistic1 | -0.156** (0.023) |
| optimistic2 | -0.117** (0.019) |
| optimistic3 | -0.034** (0.016) |
| optimistic4 | -0.013 (0.01) |
| useful1 | -0.148** (0.019) |
| useful2 | -0.114** (0.018) |
| useful3 | -0.043** (0.015) |
| useful4 | -0.02** (0.009) |
| relaxed1 | -0.153** (0.022) |

| | |
|------------------|------------------------|
| relaxed2 | -0.112** (0.018) |
| relaxed3 | -0.041** (0.015) |
| relaxed4 | -0.002 (0.009) |
| dealingproblems1 | -0.127** (0.023) |
| dealingproblems2 | -0.103** (0.018) |
| dealingproblems3 | -0.054** (0.015) |
| dealingproblems4 | -0.02** (0.009) |
| thinkingclearly1 | -0.156** (0.022) |
| thinkingclearly2 | -0.094** (0.017) |
| thinkingclearly3 | -0.021 (0.014) |
| thinkingclearly4 | 0.008 (0.01) |
| closetopeople1 | -0.205** (0.021) |
| closetopeople2 | -0.164** (0.019) |
| closetopeople3 | -0.068** (0.014) |
| closetopeople4 | -0.023** (0.01) |
| makeupownmind1 | -0.145** (0.022) |
| makeupownmind2 | -0.081** (0.018) |
| makeupownmind3 | -0.055** (0.015) |

| | |
|----------------|--------------------|
| makeupownmind4 | -0.005 (0.009) |
| constant | 0.025** (0.005) |

Number of statistically significant main effects parameters at 5% level 23

Number of statistically significant main effects parameters at 10% level 23

Notes: **significant at 5%

Robust standard errors are presented in parentheses.

Potentially logical inconsistent coefficients are highlighted in bold.

As stated in the method section, the pooled coefficients were derived from the weighted average of the C-TTO and rescaled DCE coefficients. The weights favoured the coefficients with lower standard errors (i.e. lower uncertainties), to ensure the generation of more reliable pooled coefficients. For example, as an illustration, the coefficients for the attribute levels “optimistic1” and “useful1” in Model 3 are calculated as follows:

$$\text{optimistic1} = -0.153 * \left[\frac{\frac{1}{(0.0363)^2}}{\frac{1}{(0.0363)^2} + \frac{1}{(0.0288)^2}} \right] + -0.158 * \left[\frac{\frac{1}{(0.0288)^2}}{\frac{1}{(0.0363)^2} + \frac{1}{(0.0288)^2}} \right] = -0.156$$

$$\text{useful1} = -0.144 * \left[\frac{\frac{1}{(0.0262)^2}}{\frac{1}{(0.0262)^2} + \frac{1}{(0.0279)^2}} \right] + -0.152 * \left[\frac{\frac{1}{(0.0279)^2}}{\frac{1}{(0.0262)^2} + \frac{1}{(0.0279)^2}} \right] = -0.148$$

The C-TTO and rescaled DCE coefficients for the attribute levels were extracted from the main effect models (i.e. Model 1A and Model 2A). Given the two DCE rescaling methods, coefficients for the anchoring method were selected as the rescaled DCE coefficients, as the standard errors for the anchored coefficients were already available in Table 28. For the attribute level “optimistic1”, the C-TTO standard error (0.0363) was higher than the rescaled DCE standard error (0.0288). As a result, the DCE rescaled coefficient (-0.158) with relatively lower uncertainty received a higher weight. The pooled coefficient (-0.156)

inclined towards the rescaled DCE coefficient. For the attribute level “useful1”, the C-TTO standard error (0.0262) was slightly lower than the rescaled DCE standard error (0.0279). As the two standard errors were very close, the pooled coefficient (-0.148) was roughly the mean of the C-TTO and DCE rescaled coefficients.

The pooled standard error for the attribute levels was derived under the assumption that the C-TTO and DCE rescaled coefficients were independent and normally distributed. The pooled standard error was simply the square root of the summation of weight-average C-TTO and DCE rescaled variances. For example, the standard errors for the attribute levels “optimistic1” and “useful1” are calculated as follows:

$$\text{optimistic1} = \sqrt{\left[0.0363 * \frac{\frac{1}{(0.0363)^2}}{\frac{1}{(0.0363)^2} + \frac{1}{(0.0288)^2}}\right]^2 + \left[0.0288 * \frac{\frac{1}{(0.0288)^2}}{\frac{1}{(0.0363)^2} + \frac{1}{(0.0288)^2}}\right]^2} = 0.0226$$

$$\text{useful1} = \sqrt{\left[0.0262 * \frac{\frac{1}{(0.0262)^2}}{\frac{1}{(0.0262)^2} + \frac{1}{(0.0279)^2}}\right]^2 + \left[0.0279 * \frac{\frac{1}{(0.0279)^2}}{\frac{1}{(0.0262)^2} + \frac{1}{(0.0279)^2}}\right]^2} = 0.0191$$

In addition, there was only one logically inconsistent coefficient (thinkingclearly4) in Model 3. There were 23 statistically significant coefficients at the 5% level. The summation of the absolute values of the level coefficients for each attribute is reported in Table 35. The results suggest that the ranking of the attributes based on this summation method was ordered as: closetopeople (the most important attribute), useful, optimistic, relaxed, dealingproblems, makeupownmind, and thinkingclearly (the least important attribute). Among the coefficients of all attribute levels, closetopeople1 (I’ve been none of the time feeling close to other people) received the highest weight in absolute value (0.205), indicating the largest change from the base level 5. The attribute level with the least weight in absolute value was relaxed4 (I’ve been often feeling relaxed) (0.00226), indicating the smallest change from base level 5.

Table 35: The total weight for each attribute in Model 3

| Rank | Attribute | Total weight in absolute value |
|------|---|--------------------------------------|
| 1 | I’ve been feeling close to other people | 0.205 + 0.164 + 0.068 + 0.023 = 0.46 |
| 2 | I’ve been feeling useful | 0.148 + 0.114 + 0.043 + 0.02 = 0.325 |

| | | |
|---|--|---|
| 3 | I've been feeling optimistic about the future | $0.156 + 0.117 + 0.034 + 0.013 = 0.32$ |
| 4 | I've been feeling relaxed | $0.153 + 0.112 + 0.041 + 0.002 = 0.309$ |
| 5 | I've been dealing with problems well | $0.127 + 0.103 + 0.054 + 0.02 = 0.304$ |
| 6 | I've been able to make up my own mind about things | $0.145 + 0.081 + 0.055 + 0.005 = 0.286$ |
| 7 | I've been thinking clearly | $0.156 + 0.094 + 0.021 + 0.008 = 0.279$ |

6.3.4. Sensitivity analysis

The result of the sensitivity analysis for the C-TTO main effects model is labelled as Model 4 in Appendix 25. Without excluding the value of the flagged states, it was interesting to observe that the model performance was roughly similar to that of Model 1A. Also, the variable optimistic₃, which was statistically insignificant in Model 1A, became statistically significant in this model. However, compared to Model 1A, the AIC and BIC statistics were slightly higher in Model 4 (Model 1A: AIC = 2326.403, BIC = 2654.168; Model 4: AIC = 2450.295, BIC = 2781.979). The general message was that there was no obvious difference in the model fit with and without the exclusion of flagged states identified by the participants in this study. Nevertheless, the Model 1A was still preferred to Model 4, as it was not sensible to include the values of problematic states with incorrect rank ordering claimed by the participants in the Feedback Module.

The result of the sensitivity analysis for the DCE main effects model is labelled as Model 5 in Appendix 26. After excluding potentially strategic answers, the difference in model performance between Model 5 and the Model 2A was very small. Compared to Model 2A, Model 5 possessed lower AIC and BIC statistics (Model 2A: AIC = 2849.381, BIC = 3035.273; Model 5: AIC = 2787.113, BIC = 2972.484). However, the number of statistically significant coefficients at the 5% level decreased from 21 in Model 2A to 19 in Model 5. Given that there was no large difference in model performance between these two models, Model 2A remains preferred to Model 5 as there was no evidence to justify the participants' disengagement in providing these responses under the supervision of the interviewer on task completions.

6.3.5. Comparison of valuation sets

The coefficients for the attribute levels generated by Model 1A (C-TTO main effect model) in Table 25 were comparable to those for Model 2A (DCE main effect model based on anchoring) in Table 28 as they lay on the same utility scale. The number of logically inconsistent coefficients in Model 2A was much lower than those in Model 1A. The two positive inconsistent coefficients in Model 2A belonged to level 4 of two attributes (relaxed4 and thinkingclearly4) whereas the five inconsistent coefficients in Model 1A belonged to level 2 and level 4 of five attributes (relaxed2, dealingproblems2, optimistic4, useful4, and closetopeople4). Also, the number of statistically significant parameters estimated at either the 5% or 10% level was slightly higher in Model 2A. Model 3 (IVW hybrid model) processed the highest number of statistically significant parameters estimated at either the 5% or 10% level. The number of logically inconsistent coefficients in Model 3 was also the lowest. Only one inconsistent coefficient (thinkingclearly4) was identified in this hybrid model of combining both C-TTO and DCE data.

Concerning estimation precision, the rescaled standard errors surrounding the coefficients in Model 2A were in general smaller than those in Model 1A. However, these comparative advantages of Model 2A over Model 1A did not necessarily mean that the DCE valuation technique was preferred to the C-TTO technique in modelling individual preferences towards mental well-being. Arguably, it was observed that the performance of the C-TTO model improved after adding an individual covariate in Model 1B, even though it could not be compared directly to Model 2A due to different sets of explanatory variables. Compared to Model 1A and Model 2A, the pooled standard errors surrounding all coefficients in Model 3 were the lowest.

Additionally, the result of Model 2A based on mapping DCE values onto the C-TTO comparable scale was also investigated. The utility value of the highest mental well-being 5555555 generated by mapping (i.e. 0.954) was not identical to that generated by Model 1A, Model 2A based on rescaling, and Model 3 (i.e. 1). The value of level utility change from level 5 for each attribute generated by the mapping method therefore could not be compared directly to the level coefficients generated by the C-TTO model, the DCE rescaling model based on anchoring, and the IVW hybrid model. Nevertheless, the pattern of the utility change based on the mapping method could still be explored. As with the anchoring method, there were two logically inconsistent values (relaxed4 and

thinkingclearly4) identified by the mapping method. In short, the rescaled DCE main effects model tended to perform relatively better than the C-TTO main effects model.

Table **36** below summarises the ranking of attributes based on summing the level coefficients and the attributes ranked from the largest disutility of attribute level 1 to the smallest disutility of attribute level 1. For the within-model comparison, it was noted that the two ranking methods generated slightly different attribute orders for the C-TTO and the IVW hybrid models, but the attribute orders for the two rescaled DCE models were roughly the same. When comparing within each ranking method, the ranking of attributes was different across the C-TTO, DCE and the IVW hybrid models, but they were the same between the two rescaled DCE models. Also, the largest and smallest level change from the base level were different across the C-TTO and the IVW hybrid model, and were identical between the two rescaled DCE models.

Table 36: Ranking of attributes and the largest and smallest level change from level 5

| | Model 1A - Heteroskedastic Tobit model for the C-TTO data | Model 2A - Conditional Logit model for the DCE data based on anchoring | Model 2A - Conditional Logit model for the DCE data based on mapping | Model 3 - IVW hybrid model for both C-TTO and DCE data |
|--|---|---|---|--|
| Ranking of attributes based on summing the absolute values of the level coefficients for each attribute | relaxed → optimistic → closetopeople → makeupownmind → thinkingclearly → useful → dealingproblems | closetopeople → useful → dealingproblems → optimistic → relaxed → thinkingclearly → makeupownmind | closetopeople → useful → dealingproblem → optimistic → relaxed → thinkingclearly → makeupownmind | closetopeople → useful → optimistic → relaxed → dealingproblems → makeupownmind → thinkingclearly |
| Attributes ranked from the largest disutility of attribute level 1 to the smallest disutility of attribute level 1 | thinkingclearly → closetopeople → relaxed → makeupownmind → optimistic → useful → dealingproblems | closetopeople → optimistic → useful → dealingproblems → relaxed → thinkingclearly → makeupownmind | closetopeople → optimistic → useful → dealingproblems → relaxed → thinkingclearly → makeupownmind | closetopeople → optimistic → thinkingclearly → relaxed → useful → makeupownmind → dealingproblems |

| | | | | |
|---|------------------|----------------|----------------|----------------|
| Largest utility change from the base level 5 | thinkingclearly1 | closetopeople1 | closetopeople1 | closetopeople1 |
| Smallest utility change from the base level 5 | closetopeople4 | relaxed4 | relaxed4 | relaxed4 |

In addition to investigating the utility change or coefficients of the attribute levels between the four methods (i.e. the C-TTO method, the DCE method based on anchoring, the DCE method based on mapping, and the IVW hybrid method), the performance of models could also be compared in terms of the overall valuation sets generated by these four different sources.

The utility values for some selected states calculated by Model 1A, Model 2A and Model 3 are presented in Table 37 below:

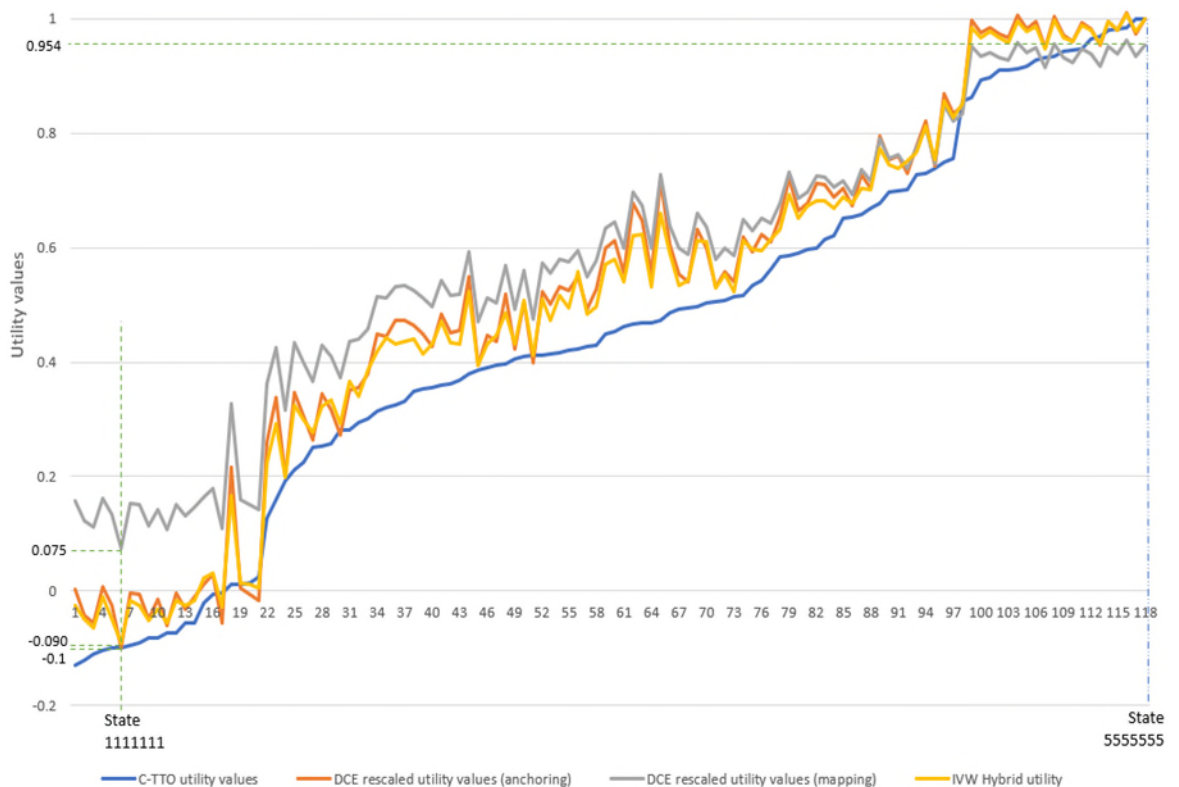
Table 37: Examples of utility calculation for particular states

| State | Model 1A - Heteroskedastic Tobit model for the C-TTO data | Model 2A - Conditional Logit model for the DCE data based on anchoring | Model 2A - Conditional Logit model for the DCE data based on mapping | Model 3 - IVW hybrid model for both C-TTO and DCE data |
|---------|---|---|--|--|
| 5555555 | 1 | 1 | $0 * 0.084 + 0.954 =$ 0.954 | 1 |
| 1111111 | $1 - 0.153 - 0.144 - 0.168 - 0.097 - 0.204 - 0.169 - 0.165 =$ -0.1 | $1 - 0.158 - 0.152 - 0.144 - 0.146 - 0.133 - 0.235 - 0.132 =$ -0.1 | $-10.46 * 0.084 + 0.954 =$ 0.075 | $1 - 0.156 - 0.148 - 0.153 - 0.127 - 0.156 - 0.205 - 0.145 =$ -0.090 |
| 3333333 | $1 - 0.058 - 0.032 - 0.089 - 0.069 - 0.046 - 0.073 - 0.082 =$ 0.551 | $1 - 0.028 - 0.046 - 0.033 - 0.05 - 0.018 - 0.067 - 0.049 =$ 0.71 | $-2.759 * 0.084 + 0.954 =$ 0.722 | $1 - 0.034 - 0.043 - 0.041 - 0.054 - 0.021 - 0.068 - 0.055 =$ 0.684 |
| 4554545 | $1 - 0.072 - 0 - 0 - 0.018 - 0 + 0.0002269 - 0 =$ 0.91 | $1 - 0.006 - 0 - 0 - 0.021 - 0 - 0.026 - 0 =$ 0.947 | $-0.504 * 0.084 + 0.954 =$ 0.912 | $1 - 0.013 - 0 - 0 - 0.02 - 0 - 0 - 0.023 - 0 =$ 0.944 |
| 2111131 | $1 - 0.135 - 0.144 - 0.168 - 0.097 - 0.204 - 0.073 - 0.165 =$ 0.014 | $1 - 0.108 - 0.152 - 0.144 - 0.146 - 0.133 - 0.067 - 0.132 =$ 0.119 | $-8.382 * 0.084 + 0.954 =$ 0.25 | $1 - 0.117 - 0.148 - 0.153 - 0.127 - 0.156 - 0.068 - 0.145 =$ 0.086 |

| | | | | |
|---------|--|---|---|--|
| 3313432 | $1 - 0.058 - 0.032 - 0.168 - 0.069 - 0.016 - 0.073 - 0.121 = \mathbf{0.463}$ | $1 - 0.028 - 0.046 - 0.144 - 0.05 + 0.011 - 0.067 - 0.063 = \mathbf{0.614}$ | $-3.671 * 0.084 + 0.954 = \mathbf{0.646}$ | $1 - 0.034 - 0.043 - 0.153 - 0.054 + 0.008 - 0.068 - 0.081 = \mathbf{0.575}$ |
|---------|--|---|---|--|

Features of these three valuation sets generated by the C-TTO and DCE rescaling models could also be investigated through a graphical plot, as shown in Figure 12 below:

Figure 12: The utility values generated by the C-TTO model, two DCE rescaling models, and the IVW hybrid model, against the SWEMWBS states ordered by the C-TTO utility values



This figure was produced by firstly sorting the 78,125 SWEMWBS states based on the level-sum scores in an ascending order. After arranging the 78,125 states from a state with the lowest level-sum score (1111111) to a state with a highest level-sum score (5555555), the horizontal axis represents every 1000th SWEMWBS states (1st, 1001st, 2001st,... . . . , 76001st, 77001st, 78001st) ordered by the C-TTO utility values. The first 20th SWEMWBS states (1st - 20th) and the last 20th SWEMWBS states (78106th – 78125th) are also included in this figure, with a view to investigating the pattern of utility values surrounding the

states with the lowest-end and highest-end level-sum scores. The closer to the left-hand side of the horizontal axis, the lower would be the C-TTO utility values of states. The closer to the right-hand side of the horizontal axis, the higher would be the C-TTO utility values of states.

The blue line represents the C-TTO utility values of the SWEMWBS states generated by the Model 1A. The dark orange line represents the DCE rescaled utility values based on anchoring for the SWEMWBS states generated by the Model 2A. The grey line represents the DCE rescaled utility values based on mapping for the SWEMWBS states generated by the Model 2A. The light orange line represents the IVW hybrid utility values generated by the Model 3.

In general, higher (lower) utility values were found for states with higher (lower) levels of mental well-being. For the valuation set generated by the C-TTO model, it was interesting that the lowest mental well-being state 1111111 did not receive the lowest utility value (-0.1). The lowest utility value calculated by this model was -0.13 for states 1122111 and 1122121. The highest utility value was 1.000227 for state 5555545, which was roughly the same as the utility value of 1 for state 5555555. For the valuation set generated by the DCE rescaled model based on anchoring, the lowest utility value was -0.1 for state 1111111. Remarkably, there were nine states (4545454, 5545454, 4555454, 5555454, 4545455, 5545455, 4555455, 5555455 and 5545555) with utility values greater than 1, indicating a utility increment from the highest mental well-being state 5555555. Among them, the highest utility value was 1.014 for state 5545455. The DCE rescaled model based on mapping was the only valuation set without any negative utility value and all utility values fell between 0 and 1. The lowest utility value was 0.075 for state 1111111, whereas the highest utility value was 0.965 for state 5545455. For the valuation set generated by the IVW hybrid model, the lowest utility value was -0.09 for state 1111111. There were four states (5545454, 5555454, 5545455 and 5555455) with utility values greater than 1.

When exploring the pattern of the three valuation sets together, it was realised that the utility values generated by the rescaled DCE model based on anchoring and mapping, and the IVW hybrid model were generally higher than that of the C-TTO model. The utility values generated by the mapping method were generally higher than those of the anchoring method and the IVW hybrid method for the range of low and intermediate

mental well-being states, but the difference diminished with increasing level of mental well-being. The utility values based on anchoring and IVW hybrid methods for some states were even higher than those generated by the mapping method at the high-end of mental well-being. Although the valuation sets generated by these two DCE rescaling methods were different to some extent, they correlated perfectly to each other at the same direction, as indicated by a correlation coefficient of 1. The IVW hybrid valuation set and either of the DCE rescaling valuation sets were also highly correlated, as indicated by the correlation coefficient of 0.992. The correlation coefficient between the IVW hybrid valuation set and the C-TTO valuation set was 0.957. The correlation coefficient between the C-TTO valuation set and either of the DCE rescaling valuation sets was also high (0.921). Furthermore, the difference between the two rescaling methods could also be compared by the MAD and RMSD statistics, as summarised in Table 38 below:

Table 38: MAD and RMSD performance of the DCE rescaling methods

| | Anchoring | Mapping |
|--|------------------|----------------|
| MAD | 0.067 | 0.05 |
| Number of predictions/states with rescaled utility value > 0.05 from observed mean C-TTO | 33 | 29 |
| Number of predictions/states with rescaled utility value > 0.1 from observed mean C-TTO | 13 | 10 |
| Number of predictions/states with rescaled utility value > 0.2 from observed mean C-TTO | 2 | 0 |
| RMSD | 0.375 | 0.367 |

In terms of exploring the predictive ability to the observed C-TTO values, the performance of the two rescaling methods was similar. Compared to the anchoring method, the MAD and RMSD of the mapping method were only slightly lower. Also, the overall deviation between the rescaled utility values and the observed mean C-TTO values for the 64 SWEMWBS states tended to be lower. There was no prediction with rescaled utility greater than 0.2 from the observed mean C-TTO.

6.4. Discussion

This chapter applies the Tobit heteroskedastic model, the conditional logit model and the IVW hybrid model to model the C-TTO and DCE responses provided by the 225 participants respectively. The results produced the first valuation sets for a mental well-being measure, which can be used for cost-utility analyses of interventions that generate well-being improvements detectable by the SWEMWBS. This section summarises the C-TTO and DCE modelling results and explores their key similarities and differences.

When exploring the performance of main effects models, the unscaled DCE model (Model 2A) was better than the C-TTO model (Model 1A) in terms of fewer potentially logical inconsistent coefficients and more statistically significant coefficients for the attribute levels. However, the result of Model 1A was obviously subject to the problem of omitted variable bias, as the explanatory power in Model 1B was enhanced after adding the covariates related to the SWEMWBS score, gender, ethnicities and education level. The decrease in the number of potentially inconsistent coefficients with smaller standard errors, the improved fitness of model identified by the decreased AIC and BIC statistics, and the statistically significant chi-squared value for the Wald test also supported the promising performance of Model 1B. As the effect of including an individual-specific explanatory variable could not be tested in the conditional logit model, this constraint limited its ability to accommodate the presence of deterministic heterogeneity to the inclusion of interaction terms only (Greene, 2003). However, it should be noted that the DCE model specification was very flexible in terms of accommodating taste heterogeneity (de Bekker-Grob *et al.*, 2012). In addition to the basic conditional logit framework, which was adopted for the purpose of testing the modelling of preliminary valuation sets in this study with small sample size, advanced choice models such as mixed logit model and latent class model could be explored in the future to model random heterogeneity.

Unlike the C-TTO model, another main difference of the DCE model was that the generated latent utility values were not bounded within a 0-1 scale to be used in economic evaluation. To allow the estimation of MWALYs based on the DCE modelling result, the C-TTO data was relied on to rescale DCE utility values onto a C-TTO comparable scale by two methods: anchoring to the lowest mental well-being state of the C-TTO and mapping DCE values onto C-TTO values. Similar to the result reported by Rowen *et al.* (2015), this study found that the utility values generated by the mapping method

performed slightly better than the anchoring method in the area of predicting the observed C-TTO values for the 64 states. However, the comparison of predictive ability to the observed C-TTO values measured by the MAD or RMSD across the two methods in this study was rough indicators, as it was argued that the mean observed value was no longer the best indication of central tendency because the C-TTO values were censored (Feng *et al.*, 2018). Also, there were different constraints for both rescaling methods. For the anchoring method, the DCE latent utility value for the lowest mental well-being state was anchored to the value of the lowest mental well-being state for the C-TTO. As this rescaling method required consistency in the number of explanatory variables between the C-TTO model and the DCE model, the DCE rescaled coefficients could unfortunately only be derived from the Model 1A, which was a less preferred model than the Model 1B. This rescaling constraint was not applied to the mapping method, as the dependent variable of its specification was the observed mean C-TTO value of the 64 states based on the responses by the 225 participants. This was obviously a more flexible method without the need to rely on the C-TTO modelling result in Table 25. Nevertheless, there were shortcomings relative to the anchoring method. The anchoring method was a user-friendly approach to users or policy makers, as the rescaled coefficients associated with their standard errors representing the utility change from the reference level 5 for all attribute levels could be derived easily without the use of modelling techniques. Resource allocation decisions could be guided by conveniently referring to the relative importance of each attribute level to the general population. However, for the mapping method, the magnitude of utility change from the reference level was not straightforward as the utility value for the highest mental well-being state was not 1. Future research will also be required to model the reliability and efficiency of these utility changes through the estimations of confidence intervals and standard errors. There were independent DCE rescaling methods without the need to depend on the data from C-TTO or other valuation techniques. For example, Norman *et al.* (2014) included the survival duration attribute in addition to the SF-6D dimensions within the DCE choice sets to explore the trade-off between quality of life and life expectancy. A ratio of marginal utilities approach representing the ratio of the change of utility given a unit change in the duration of one alternative over another alternative was used to generate the QALY weights. Stolk *et al.* (2010) attempted to anchor the DCE values using the dummy coefficient for dead. Respondents were asked to compare each of the two EQ-5D health profiles within the

DCE pair to a choice of dead. Based on the rank ordering result of the two states in each pair and the dead state, a rank-ordered logit model including the “dead” parameter was used to generate the rescaled values. Despite the existence of these rescaling methods, there was a lack of research investigating the way to convert the DCE values into a 0-1 scale without relying on the data from other valuation techniques when there was no dead or duration attribute available in the choice tasks.

Moreover, similar to the sensitivity analysis result of the Danish EQ-5D-5L value set (Jensen *et al.*, 2021), substantial difference of model performance with and without the exclusion of the flagged C-TTO values in the Feedback Module was not observed in Model 4. Arguably, the similar results between Model 1A and Model 4 could be partly due to the small number of flagged answers, as only 147 (6.53%) out of the 2250 responses were deleted in Model 1A. Notwithstanding, as the results from other country-specific EQ-5D-5L value sets showed an improvement of goodness of fit and smaller number of statistically insignificant parameters after excluding the flagged states (Ferreira *et al.*, 2019; Wong *et al.*, 2018), it will be valuable to analyse the role of the Feedback Module on modelling implications under a larger sample size. Furthermore, although some potentially strategic DCE responses were identified in this study, these responses were not deleted because of insufficient evidence to prove the insincerity of these answers. The presence of an interviewer was important to ensure the quality of these responses as the validity of these responses could be judged by the supervision of participants’ engagement behaviour (e.g. the speed or mood of completion).

In addition to separately modelling the C-TTO and DCE responses, this chapter also explored the IVW hybrid approach for modelling both C-TTO and DCE responses (Model 3). I am not aware of the application of this approach within the previous health state preference elicitation studies. This approach applied the C-TTO standard errors and rescaled DCE standard errors for deriving the weights for calculating the pooled coefficients and standard errors. Due to the easily accessible standard errors of coefficients for the anchoring method, the standard errors of coefficients derived from the anchoring method rather than the mapping method was used in the calculation of DCE weights for the attribute levels. As this method attached more weight to parameters with lower standard errors, it was sensible to discover that the correlation between the IVW hybrid valuation set and the DCE anchoring valuation set (0.992) was higher than the correlation

between the IVW hybrid valuation set and the C-TTO valuation set (0.957). Model 3 was better than Model 1A and Model 2A based on anchoring in terms of the least number of potentially logical inconsistent coefficients, the highest number of statistically significant coefficients, and the lowest standard errors surrounding the coefficients of all attribute levels. Obviously, comparing across all modelling approaches in this chapter, this IVW hybrid approach was a potential candidate for deriving the most optimal valuation set. It offered a balanced perspective in gathering the preference information elicited from two different valuation techniques.

Whilst this study focused on the application of C-TTO and DCE methods for the valuation of the SWEMWBS, future research could also test the validity of other valuation techniques. Recently, there has been a development of valuation approaches other than traditional valuation techniques. The main feature of these approaches is that the social value set can be conveniently derived from the personal utility function of each participant without the need to apply modelling techniques. For example, Devlin *et al.* (2019) aimed to test the elicitation of personal utility functions by directly asking participants in England the relative importance of the EQ-5D dimensions, levels and the associated interactions. The valuation procedure began with participants reporting their own EQ-5D and EQ-VAS ratings, followed by ranking of the five dimensions. The method of swing weighting was then applied to allow participants to rate the importance of dimensions based on the improvement from the worst level to the best level. After that, they were entered the level rating task. According to the results from the dimension rating and level rating tasks, they were given different pairs of choice tasks and they were asked to decide the preferred option between two health states without duration. Next, several questions comparing immediate death to living in a particular health state for 10 years were asked to derive the location of dead for each person. Lastly, interactions between levels were explored. Additionally, Sullivan *et al.* (2020) also developed a new online tool using “1000minds” software to derive personal and social EQ-5D-5L value sets in New Zealand. Participants were asked several adaptive DCE questions, selecting between living in two health states with two EQ-5D-5L dimensions each within a given life span, followed by binary search questions to identify the diving threshold when moving from better-than-dead to worse-than-dead states. Personal weights for each dimension level across all participants were averaged to generate a social value set. This form of DCE questioning with the inclusion of fewer number of attributes in the context of choice pair comparison could be an obvious

advantage for the valuation of SWEMWBS. The problem of overwhelming information, as identified by the results of the think-aloud study in Chapter 4, caused by comparing seven-to-seven attributes with different combinations of levels, could be relieved.

The main modelling limitation of this study was the small sample size. Incentive problem was one of the main hurdles in sample recruitment. Due to limited budget, the money voucher was not given to all participants and only each of the randomly drawn 10 winners received this financial reward. As a result, some potential participants were missed after realising that there were no guaranteed money reward or other non-financial rewards after their participations. As only 225 participants were interviewed, the statistical power of the C-TTO and DCE models was limited. It was therefore not surprising to realise that several statistically insignificant coefficients for the attribute levels were identified. Nevertheless, given that the idea of this study was to test the validity of the valuation protocol, the valuation sets generated in this chapter were preliminary. Moreover, the decision around the types of covariates included in the modelling analysis was informed by the result of the qualitative phase and the amount of demographic information collected during the interviews in the quantitative phase. Other than investigating the interaction effect between participants' ages and the attribute "I've been feeling useful", future research could explore the interactions between the attribute levels and other demographic variables collected from the interviews.

6.5. Conclusion

This chapter provides the first attempt to model the preliminary value sets for the SWEMWBS. This chapter compared the nature and difference of the C-TTO, DCE and the IVW hybrid modelling results and facilitated an understanding of the relative importance of different mental well-being attributes. The modelling results represent the first of their kind valuation sets of a generic mental well-being measure. These results can now be used to explore policy implications in economic evaluations. They also have the potential to inform modelling implications of the future national valuation studies of SWEMWBS. Even though the IVW hybrid model offered a statistical way of combining both C-TTO and DCE data optimally, the best tariff set should ultimately be informed by modelling results generated from surveys of larger samples of the general population.

Chapter 7: Discussion and conclusion

7.1. Introduction

This thesis explores the derivation of preliminary preference-based valuation sets for the SWEMWBS. In this chapter, a summary and discussion of findings from different stages of this PhD is provided to integrate the answers to the four main research questions as mentioned in Chapter 1:

- Do any existing preference-based measurement approaches and instruments value mental well-being?
- Are there any mental well-being measures that can be used to develop a preference-based tariff?
- What is the best choice of instrument for the elicitation of a preference-based tariff to allow the calculation of MWALY?
- What is the appropriate valuation protocol for the valuation of mental well-being state?

Also, the application and role of the derived valuation sets on economic evaluations and their policy implications on decision making will be discussed. Finally, contributions of this PhD research and a discussion of limitations and future research agenda will be provided.

7.2. Summary and discussion of main results to the research questions

7.2.1. Do any existing preference-based measurement approaches and instruments value mental well-being?

Chapter 2 firstly discussed this research question by reviewing the theoretical concepts of different preference-based measurement approaches and comparing the existing generic preference-based instruments. The MAU instruments (QWB-SA, EQ-5D-5L, HUI3, SF-6D, 15D, AQoL-8D and ReQoL) and capability measures (ICECAP-A, ICECAP-O and ASCOT) were compared in terms of their coverages on physical health and mental health dimensions. The results showed that although there are preference-based instruments valuing mental well-being, the extent or focus of their coverages of mental well-being is limited. The attributes or dimensions of most of the existing MAU instruments focused mainly on the constructs related to physical health, without capturing broader aspects of

mental health. It was discovered that AQoL-8D and ReQoL could be more sensitive in capturing the value of mental well-being. However, there was a lack of published evidence supporting the performance of AQoL-8D in the economic evaluation of mental well-being interventions, due to its significant focus on aspects of negative mental health. Also, the ReQoL was developed for use in mental health service users, but not the general population. The use of this questionnaire in general population might exhibit ceiling effects. The inclusion of a physical health question in the questionnaire also reveals that ReQoL is not a pure mental well-being instrument. Finally, the capability measures focus on the theoretical concepts of functioning and capability (Coast *et al.*, 2008c; Karimi *et al.*, 2016). The attributes covered within the questionnaires mainly assess the individual's ability to achieve combinations of functioning. The assessment of mental well-being is not the core role of these measures.

7.2.2. Are there any mental well-being measures that can be used to develop a preference-based tariff?

Existing non-preference-based mental well-being instruments (WEMWBS/SWEMWBS, WHO-5, MHC-SF and SEHS) were compared in terms of their coverages on hedonic and eudaimonic well-being. Those instruments with a pure focus on mental well-being (i.e. positive mental health) concepts and 100% positively worded items were included in the comparative analysis. Among them, the SEHS and SWEMWBS have the largest coverage proportion on functioning, as 80-99% of the 36 items in SEHS and 7 items in SWEMWBS are related to the assessment of eudaimonic well-being (Rose *et al.*, 2017). The WHO-5 has the largest coverage proportion on feeling, as 80-99% of the 5 items are related to the assessment of hedonic well-being. Different from WEMWBS/SWEMWBS, WHO-5 and MHC-SF, the SEHS is the only measure with the inclusion of school-based items, as it was developed for use by adolescent populations. All these measures could be potential candidates for the elicitation of a preference-based tariff to estimate the MWALY.

7.2.3. What is the best choice of instrument for the elicitation of a preference-based tariff to allow the calculation of MWALYs?

Among the choices of non-preference-based mental well-being measures, SWEMWBS was identified as the best candidate. As the evaluation of well-being is not only restricted to understand the feelings and moods of an individual, eudaimonic view of well-being plays a more comprehensive role in investigating higher level of well-being in terms of

achieving the goal of self-actualization. In this context, compared to the WHO-5, the items in SWEMWBS are better in covering different aspects of eudaimonic well-being. The generic nature of SHES is limited by its coverage of several items related to the individual's belief in school teachers. Unlike SWEMWBS, it is suitable for the completion of students with a specific age group only. MHC-SF is not as popular as SWEMWBS regarding the published evidence of psychometric validation in the UK. Complementing this comparative result with other unique strengths of SWEMWBS, a summary of the reasons for regarding SWEMWBS as the best choice of instrument is as follows:

- It has a unique and broad coverage of eudaimonic well-being (Rose *et al.*, 2017).
- It has been widely applying in different sectors of the economy such as health, business and education (Shah & Stewart-Brown, 2017).
- A high rating on the WEMWBS was received by mental health care service users (Crawford *et al.*, 2011).
- It is used by policy makers in Scotland, Wales and England for monitoring population mental well-being (Diana Bardsley *et al.*, 2017; Parkinson, 2007; Scottish Government, 2018).
- Its popularity is further supported by the increasing annual trends of usage, increasing number of registrations, increasing number of publications (Shah *et al.* (2017b)).
- It has been widely validated in different population groups with robust psychometric properties, in terms of responsiveness to mental health interventions, normally distributed in most populations, uni-dimensionality, internal consistency, construct validity, convergent validity, concurrent validity, test-retest reliability, face validity, discriminant validity, and the absence of floor and ceiling effects for the response levels (Bass *et al.*, 2016; Haver *et al.*, 2015; Koushede *et al.*, 2019; Ng Fat *et al.*, 2017; Ng *et al.*, 2014; Rogers *et al.*, 2018; Vaingankar *et al.*, 2017).

7.2.4. What is the appropriate valuation protocol for the valuation of mental well-being state?

Chapters 3 - 6 answered this question by reviewing the stages and methods of preference elicitation, followed by conducting two testing phases (i.e. a qualitative phase and a quantitative phase) to explore the validity of the proposed valuation protocol for the SWEMWBS. In this thesis, the C-TTO and DCE were adopted for the valuation of

SWEMWBS. The EQ-VT 2.1 protocol invented by the EuroQol Group (Stolk *et al.*, 2019) was closely followed throughout the two testing phases, with some modifications concerning the features and layout of tasks to adapt the difference between health state valuation and mental well-being state valuation. The proposed valuation protocol was firstly tested in the qualitative phase in Chapter 4 by the technique of cognitive interview among 14 interviewees. Through understanding participants' cognitive process of completion, the valuation protocol was modified and further tested within a larger sample of 225 participants in the quantitative phase. Psychometric validity of the valuation techniques and the modelling of preliminary value sets were analysed in Chapters 5 - 6 to explore the robustness of valuation protocol in reflecting individual preferences about mental well-being.

Whilst participants in the qualitative phase found the valuation tasks generally manageable, six broad themes emerged to explain and optimise the response to the tasks. 1) Format and structure: attention should be paid to the design of practice examples, instructions, and lay-out, to suit people from different backgrounds. 2) Items and levels: underlying relationships were likely across different combinations of levels of SWEMWBS items, which had modelling and valuation implications. 3) Decision strategies: participants engaged in strategies (i.e. interpretation of meaning of items, lexicographic ordering, interpretation of levels, comparison with previous tasks, consideration of personal and external factors, availability heuristic, duration of TTO states, satisficing heuristic, ignorance of identical levels of items between DCE alternatives, rejection of unimaginable state, framing effect and integration of self-written notes) to assist trade-off decisions. 4) Valuation feasibility: certain mental well-being states were difficult to imagine, compare and quantify. 5) Valuation outcome: The quality of the data was affected by participants' discriminatory ability across mental well-being states, their time trade-off decisions, and their ability to choose between forced DCE choices. 6) Reflections on mental well-being: The usefulness of these valuation tasks on reflecting personal preferences enhanced the practicality of using techniques widely used for health state valuation for valuing mental well-being.

The results of the quantitative phase showed that the application of the C-TTO and DCE valuation techniques in the valuation of SWEMWBS was feasible and practical when analysing the debriefing statements about the participants' cognitive burden of completion,

and the statistics properties of the C-TTO and DCE responses. The face validity was also confirmed by comparing the mean values of the C-TTO states and the level-sum score of the DCE states based on different levels of mental well-being. The modelling results of the C-TTO and DCE responses generated versions of valuation sets, which can be used in the cost-utility analysis of mental well-being interventions. The tariff generated by the IVW hybrid model seems to provide an optimal view of balancing both C-TTO and DCE preference information, and process desirable statistical properties in terms of fewer potentially logical inconsistent coefficients and lower standard errors. As different valuation techniques gather different perspectives of preference information, this hybrid method could be viewed as stronger in reflecting broader aspects of preferences and being less influenced by valuation bias of a single valuation technique.

Synthesising the results from both qualitative and quantitative phases, it was discovered that some of the modified elements in the valuation protocol informed by the qualitative phase caused certain impacts or reflections on the result of quantitative phase. They are stated in Table 39 below:

Table 39: The efficacy of applying the modified valuation protocol to the quantitative phase

| Issue identified in the qualitative phase | Proposed modification to the valuation protocol | Outcome in the quantitative phase |
|--|---|---|
| Inappropriate C-TTO practice examples | An alternative version of practice example related to physical health and relationship was added. | 56.44% of participants selected this alternative version. Comprehension problems were found around a few participants. |
| Confusion about the time trade-off procedure | <ul style="list-style-type: none"> ▪ More detailed explanations of the instructions. ▪ Slowing the instructing speed. ▪ Encouraging participants to raise questions. ▪ Clarification of practice states before completion. ▪ More step-by-step trade-off demonstrations. | More than 90% of participants somewhat or strongly agreed that they did not have difficulty in following and understanding the instructions of the tasks. |

| | | |
|--|--|---|
| Visual difficulty in differentiating the states within the C-TTO feedback module | Guidance to enhance the readability of the states line-by-line will be provided. | <ul style="list-style-type: none"> ▪ No significant problem was found in understanding the layout of this summary slide. ▪ Disagreement was confirmed when participants flagged problematic states. |
| Incomprehensible combinations of levels of attribute | The selection of experimental design choice sets with potential uncommonly reported states could be avoided. | Potentially uncommonly reported states identified in the qualitative phase were not identified in both C-TTO and DCE designs. |
| The existence of preference heterogeneity | Advanced modelling techniques can be applied to model deterministic and random heterogeneities. | Covariates and interaction terms were added to model the C-TTO and DCE data. |
| Promising manageability of the number of tasks | The number of tasks for each of the C-TTO and DCE parts will be increased from 8 to 10. | Participants were comfortable to complete more than 10 tasks for each of the C-TTO and DCE parts. |

Firstly, some participants in the qualitative phase raised the imagination difficulties in the C-TTO practice scenarios as job application was irrelevant to their current life experience. A more generic version of example including the imagination of mental reaction caused by physical health issue was added to enhance the participants' flexibility in choosing the most suitable version with the least imagination burden. The result in the quantitative phase showed that more than half of the participants chose this alternative version of C-TTO example, revealing the important role of having this alternative choice to participants. However, even though most participants were able to understand the examples, more comprehension problems related to the description of scenarios were identified in this example, compared to the version related to job application. It was discovered that there were more "noises" or factors guiding the mental reaction to physical health problems. It was therefore debatable and difficult to generalise the overall mental description in the scenarios. To avoid spending too much time on clarifying and explaining the practice scenarios, as stated in the discussion section in Chapter 5, the elements or tasks included in the practice examples could be refined.

Secondly, concerning the confusion about the time trade-off procedure identified in the qualitative phase, the participants in the quantitative phase were able to complete the practice tasks with less operational problems under the improved instructing style. Instant feedback on potential misunderstanding of each move was also received when demonstrating step-by-step examples. This could eliminate the accumulation of comprehension problems in the subsequent tasks.

Thirdly, several participants in the qualitative phase expressed the visual confusion in the Feedback Module. The inclusion of pictures or colours instead of displaying the plain text of all states were not considered in the quantitative phase, due to the difficulty in finding appropriate visual images and concern of further confusions caused by misinterpretation of images or mixture of colour texts. Alternatively, the description of this summary slide was improved by explaining explicitly the meaning of the position of each state with its corresponding rank ordering number. It was encouraging that no specific concerns about the readability of this slide were raised by participants in the quantitative phase, even if the structure of the Feedback Module was complicated by an increase in the number of states from 8 in the qualitative phase to 10 in this phase.

Fourthly, some participants in the qualitative phase raised the illogical combination of certain levels of attributes within the states. No constraint on specific combination of attribute levels was set within the C-TTO and DCE experimental designs in the quantitative phase to maintain statistical efficiency. However, as a way to enhance experimental realism, the selected C-TTO and DCE subsets derived from numerous iterations were checked to confirm that the potentially uncommonly reported states (i.e. D-values of 10 or above) were not included. The debriefing statistics showed that around 80% of the participants somewhat agreed or strongly agreed with the statement that “Overall, I didn’t have difficulty in imagining each of the allocated mental well-being states”. This result supported the acceptability of the imagination burden in general.

Fifthly, regarding the possibility of having different preferences across participants with different demographic backgrounds, covariates and interaction terms were included in the C-TTO and DCE model specifications in addition to the main effects parameters. Also, informative priors about the preference characteristics of the attribute levels were included in the utility specification of the DCE experimental design. The modelling results produced in Chapter 6 can inform future priors in the DCE experimental design. Different

types of covariates and interaction terms based on the participants' demographic information could also be explored in the future with the aid of more advanced modelling techniques.

Sixthly, it was realised that the manageable number of tasks completed by participants could go beyond 10 for each part of the interview. This suggested that the adjustment from 8 tasks in the qualitative phase to 10 tasks in the quantitative phase was appropriate. Future research could continue to explore the optimal number of tasks for each participant.

Apart from reflecting the effect of applying the modified protocol on the quantitative outcome, it was important to note some of the common findings identified by both the qualitative phase and quantitative phase, but not documented in the previous health preference research. The inferiority of full mental well-being was raised by a few numbers of participants (i.e. one out of 14 participants in the qualitative phase and three participants out of 225 participants in the quantitative phase) for the completion of C-TTO tasks. This issue had an important implication on the modelling of mental well-being state, as the highest mental well-being state (5555555) was not necessarily the most preferred status. The appropriateness of censoring responses with C-TTO value greater than one to an upper limit of 1 within the C-TTO model is debatable, as an EQ-5D-5L quality assurance article mentioned that it was impossible to have TTO values greater than 1 (Alava *et al.*, 2020). Censoring at 1 was inappropriate and the upper bound should be modelled as an inherent limit. However, arguably, full health (i.e. 11111 described by the EQ-5D-5L classification system) and full mental well-being (i.e. 5555555 described by the SWEMWBS classification system) are different concepts. It was impossible for the TTO value to exceed 1 when valuing the EQ-5D-5L health states, as the EQ-5D-5L descriptive system focuses extensively on assessing physical health dimensions. There should be an obvious discount to EQ-5D-5L score when individuals suffered from problems related to mobility, self-care, usual activities and pain/discomfort. It was illogical for an individual to not preferring full health, as individual won't prefer to have an illness or impairment on their daily physical health. The only mental health dimension for the EQ-5D-5L descriptive system is "anxiety/depression", which is negatively worded. It was also illogical for an individual to prefer any forms of mental illness. Considering these, no problems for all five dimensions (i.e. 11111) for the EQ-5D-5L should be the best state for an individual. In terms of the SWEMWBS descriptive system, it focuses on assessing

mental well-being (i.e. positive spectrum of mental health) of an individual. All items are positively worded. Based on the qualitative interview findings in Chapter 4, it was possible for a participant to prefer a state which was lower than full mental well-being. The reasons for not preferring “all of the time” for all SWEMWBS items (e.g. prefer “often” for some items) were that she considered maximal well-being state as a lack of challenging life experience, which was a crucial element of an exciting and balanced life. It was unhealthy to not having ups and downs in their mental life. For some items such as “all of the time feeling useful”, it was too tough for her as she sometimes preferred to enjoy herself or to be lazy instead. These reasons showed that it was completely sensible for a participant not to prefer full mental well-being (i.e. C-TTO value >1) as individuals valued the crucial elements for their mental health differently. Similar results were also discovered for a few participants in the quantitative phase. Three participants did not prefer full mental well-being for some C-TTO tasks and they preferred life B (i.e. a state lower than full mental well-being) instead at the beginning of the tasks. After a brief follow-up on their answers at that time, their justification of not preferring full mental well-being was roughly similar to the reasons raised by the participant in the qualitative phase. Given that it was possible to have C-TTO value greater than 1 for the valuation of SWEMWBS and only a small proportion of responses (17 responses, occupying only 0.76% of the total responses) exhibited this non-monotonic preference, it was legitimate to censor those responses with C-TTO values greater than one to 1 when modelling the C-TTO data in the quantitative phase. Also, it should be noted that for the QALYs, the highest possible state (i.e. full health) represents only really a ‘good’ health, i.e. what a reasonably fit person might comfortably achieve. Maximal mental well-being, however, is much stronger than this, and is unlikely to be a state reflecting normal unimpaired life. Future research could be cautious in judging the participants’ rationality, given that it is possible and sensible for a close to highest mental well-being state to be more preferable than the highest mental well-being state. Finally, the C-TTO results from the two piloting phases both showed that participants were willing to sacrifice years of life for a better mental well-being. This signified the importance of targeting mental well-being throughout this thesis.

7.3. Application and role of the valuation sets

The derived valuation sets could be used to calculate MWALYs for the cost-utility analysis of mental well-being interventions. The estimation of MWALYs and choice of tariff has important implications on recognising the benefits provided by the interventions. As an illustration, as shown in Table 40 below, consider a hypothetical example of two interventions related to yoga and teaching with the same implementation cost.

Table 40: The influence of tariff choice on the effectiveness of interventions with identical implementation costs and change in outcome score

| | Yoga | Language teaching for refugee children |
|---------------------------------|---|---|
| Costs | £2000 | £2000 |
| Baseline mean SWEMWBS score | 8 (i.e. 1121111) | 8 (i.e. 1211111) |
| Mean SWEMWBS score after 1 year | 12 (i.e. 1151211) | 12 (i.e. 1511121) |
| MWALYs gain | C-TTO: $0.238*1 = 0.238$ DCE (anchoring): $0.149*1 = 0.149$ DCE (mapping): $0.118*1 = 0.118$ IVW hybrid: $0.181*1 = 0.181$ | C-TTO: $0.117*1 = 0.117$ DCE (anchoring): $0.186*1 = 0.186$ DCE (mapping): $0.148*1 = 0.148$ IVW hybrid: $0.155*1 = 0.155$ |
| Costs per MWALY gain | C-TTO: £ 8403.36 DCE (anchoring): £13440.41 DCE (mapping): £16949.15 IVW hybrid: £11049.88 | C-TTO: £ 17094.02 DCE (anchoring): £10738.67 DCE (mapping): £13513.51 IVW hybrid: £12895.66 |

Due to limited public funding, resources should be allocated to the implementation of the most cost-effective intervention. Assume the costs of intervention and the baseline mean SWEMWBS scores for the intervention groups between the two interventions were the same. There was no difference on baseline characteristics between the intervention arm and the control arm for both interventions. After implementing both interventions for one year, it was observed that both interventions improved the mean SWEMWBS scores for the intervention arms from 8 to 12, compared to the control arms. Without recognising the MWALYs gain associated with each intervention, it seemed that both interventions were equally effective in generating a seven-point improvement in mental well-being and policy makers should be indifferent between these two interventions. However, this conclusion

is not informative. The interpretation of effectiveness based on level-sum scores neglected the component changes of the attribute levels. There were many attribute-level combination possibilities for constituting the aggregate score of 12 (e.g. 1122222, 3112113, 2111151, etc.). Obviously, different combination possibilities implied different policy implications. For example, the combination possibility of 2111151 means that the intervention was the most effective in improving the close feeling to other people for the population. This signified the importance of estimating the MWALYs, as the resource allocation decision could be targeted according to the public preferences (i.e. utilities) towards the seven attributes of the SWEMWBS. As the attributes of SWEMWBS were not perceived as equally valuable, the MWALYs gain associated with an intervention was based on which items were affected, rather than the quantity of outcome change.

Furthermore, as different valuation methods could yield different estimations of MWALYs, the choice of tariff matters a lot. Suppose the two interventions were targeting at improving different aspects of mental well-being. The yoga intervention helped the intervention arm stay calm and reduced the level of stress, compared to the control arm. It contributed a significant improvement of relaxation score from level 2 to level 5 for the attribute “I’ve been feeling relaxed” after twelve months of implementing this intervention. It also helped the intervention arm think slightly clearly than before, improving the level of the attribute “I’ve been thinking clearly” from level 1 to level 2. When calculating the MWALYs gain associated with the change in attribute-level combination from 1121111 to 1151211 caused by this intervention, different valuation sets derived in Chapter 6 led to different results. The ranking of the attribute “I’ve been feeling relaxed” based on summing the level coefficients, as discussed in

Table **36** of Chapter 6, was the 1st for the C-TTO model. This implied that the general population valued this attribute as the most important one out of all seven items. The ranking of this attribute for the DCE model based on anchoring and mapping was both at the 5th position. The ranking of this attribute for the IVW hybrid model was the 4th, which lay between the ranking of the C-TTO model and the rankings of the two DCE models. Based on the high weights contributed by the C-TTO coefficients for the levels of this attribute, it was not surprising to realise that the MWALYs gain calculated using the C-TTO utility value were the highest. The MWALYs calculated using the two DCE models and the IVW hybrid model were relatively lower, as these models contributed less weights for the improvement of relaxation attribute. The between-model impact of the attribute “I’ve been thinking clearly” among the C-TTO model, DCE models, and the IVW hybrid model was negligible, due to the similar rankings for this attribute based on the summation method and the small change in the attribute level score. Obviously, this yoga intervention achieved the lowest costs per MWALY gain estimated by the C-TTO tariff. The costs per MWALY gain calculated using the IVW hybrid tariff captured the tariff features of both C-TTO and DCE models and therefore bounded between the costs per MWALY gain for the C-TTO and DCE models.

In addition, another intervention allowed participants in the intervention arm to deliver English language teaching classes to refugee children whose mother language was not English. As teaching provided satisfactions and recognition of self-values to the intervention arm, there was a great improvement of usefulness feeling score from level 2 to level 5 for the attribute “I’ve been feeling useful” after twelve months of implementing this intervention. The intervention arm also found that this well-organised teaching programme helped them feel slightly closer to people (i.e. children) than before. The rankings of the attributes “I’ve been feeling close to other people” and “I’ve been feeling useful” for the two DCE models and IVW hybrid model were 1st and 2nd respectively, whereas the rankings based on summing the level coefficients were 3rd and 6th respectively for the C-TTO model. As the C-TTO model placed relatively less emphasis on the most improved attributes associated with this intervention, the MWALYs gain calculated using the C-TTO model was less than the DCE models and the IVW hybrid model.

The above example showed that the costs per MWALYs gain associated with these two interventions were influenced by the choices of tariff. Even if the costs and changes in

mean SWEMWBS scores were the same across these two interventions, the implementation recommendation could be different when applying different valuation sets. When adopting the C-TTO tariff, the yoga intervention was preferred or more cost-effective than the teaching intervention as the yoga intervention achieved lower costs per MWALY gain. The implementation of the yoga intervention would be more beneficial to the general population. However, when adopting the two DCE tariffs and the IVW hybrid tariff, the teaching intervention was preferred or cost-effective than the yoga intervention, as the costs per MWALY gain calculated using the C-TTO tariff for the teaching intervention were higher. The implementation of the teaching intervention would be more beneficial to the general population under this circumstance. Furthermore, it was also realised that the MWALYs gain calculated using the DCE model based on mapping was lower than those for the DCE model based on anchoring and the IVW hybrid model among the two interventions. One of the possible reasons could be that the standard deviation for the DCE tariff based on mapping (0.13) was lower than those for the DCE tariff based on anchoring (0.16) and the IVW hybrid tariff (0.16). As a result, the magnitude of utilities gain associated with the interventions calculated using the DCE tariff based on mapping was always the lowest.

Moreover, the derived valuation sets could also be applied to analyse the effectiveness of an intervention across different subgroups for the intervention arm. For instance, extending the example of the yoga intervention above, Table 41 below provides information related to the change in mean SWEMWBS score across the deprived group and the high-income group.

Table 41: Sub-group analysis for the effectiveness of an intervention with different magnitude changes in outcome score between groups

| | Yoga | | |
|---------------------------------|---|---------------------------|---------------------------|
| Costs | £2000 | | |
| Baseline mean SWEMWBS score | Low-income group = High-income group = 7 (i.e. 1111111) | | |
| Mean SWEMWBS score after 1 year | Low-income group: 14 (i.e. 2151311); High-income group: 23 (i.e. 3253433) | | |
| MWALYs gain | | Low-income group (70%) | High-income group (30%) |
| | C-TTO | $[0.344*1] * 0.7 = 0.241$ | $[0.685*1] * 0.3 = 0.206$ |

| | | | |
|----------------------|-----------------|---------------------------------|---------------------------------|
| | DCE (anchoring) | $[0.309*1] * 0.7 = 0.216$ | $[0.804*1] * 0.3 = 0.241$ |
| | DCE (mapping) | $[0.247*1] * 0.7 = 0.173$ | $[0.643*1] * 0.3 = 0.193$ |
| | IVW hybrid | $[0.326*1] * 0.7 = 0.228$ | $[0.774*1] * 0.3 = 0.232$ |
| Costs per MWALY gain | | Low-income group | High-income group |
| | C-TTO | $2000/0.241 = \text{£}8305.65$ | $2000/0.206 = \text{£}9732.36$ |
| | DCE (anchoring) | $2000/0.216 = \text{£}9253.69$ | $2000/0.241 = \text{£}8287.87$ |
| | DCE (mapping) | $2000/0.173 = \text{£}11567.38$ | $2000/0.193 = \text{£}10368.07$ |
| | IVW hybrid | $2000/0.228 = \text{£}8753.13$ | $2000/0.232 = \text{£}8615.03$ |

Assume that the intervention arm included 70% of low-income participants and 30% of high-income participants. The yoga intervention improved the mental well-being of the intervention arm from a SWEMWBS score of 7 to 14 for the low-income group and from 7 to 23 for the high-income group. The MWALYs gain associated with the intervention for the two subgroups were calculated proportionally. Even though the SWEMWBS score improvement for the high-income group was much higher than that for the low-income group, this did not necessarily mean that resources should be allocated to provide yoga intervention for the high-income group only. When using the C-TTO tariff for the calculation of costs per MWALY gain, it was more cost-effective to provide this intervention for the low-income group. The implementation of this intervention was only more cost-effective for the high-income group when applying the two DCE tariffs and the IVW hybrid tariff, as indicated by the lower costs per MWALY gain. This example showed that resource allocation decision could be targeted according to the demographic characteristics of the intervention arm. The benefits of providing the intervention to a larger proportion of people with lower improvement in outcome score (i.e. low-income group) could be higher than providing the intervention to a lower proportion of people with greater improvement in outcome score (i.e. high-income group).

Finally, consider another scenario in Table 42 below that the proportion of the two groups was distributed evenly within the intervention arm. The MWALYs gain associated with an identical magnitude improvement of the level score for the same attribute could be different when the baseline mean scores between groups were not the same.

Table 42: Sub-group analysis for the effectiveness of an intervention with identical magnitude change in level score for the same attribute between groups

| | | Yoga | |
|---------------------------------|---|---------------------------------|----------------------------------|
| Costs | £2000 | | |
| Baseline mean SWEMWBS score | Low-income group = 14 (i.e. 2222222); High-income group = 21 (i.e. 3333333) | | |
| Mean SWEMWBS score after 1 year | Low-income group: 16 (i.e. 2242222); High-income group: 23 (i.e. 3353333) | | |
| MWALYs gain | | Low-income group (50%) | High-income group (50%) |
| | C-TTO | $[0.123*1] * 0.5 = 0.062$ | $[0.089*1] * 0.5 = 0.045$ |
| | DCE (anchoring) | $[0.089*1] * 0.5 = 0.045$ | $[0.033*1] * 0.5 = 0.016$ |
| | DCE (mapping) | $[0.071*1] * 0.5 = 0.036$ | $[0.026*1] * 0.5 = 0.013$ |
| | IVW hybrid | $[0.11*1] * 0.5 = 0.055$ | $[0.041*1] * 0.5 = 0.021$ |
| Costs per MWALY gain | | Low-income group | High-income group |
| | C-TTO | $2000/0.062 = \text{£}32520.33$ | $2000/0.045 = \text{£}44943.82$ |
| | DCE (anchoring) | $2000/0.045 = \text{£}44907.16$ | $2000/0.016 = \text{£}122303.42$ |
| | DCE (mapping) | $2000/0.036 = \text{£}56338.03$ | $2000/0.013 = \text{£}153846.15$ |
| | IVW hybrid | $2000/0.055 = \text{£}36294.55$ | $2000/0.021 = \text{£}97223.18$ |

Assume the baseline mean SWEMWBS score for the high-income group was higher than that for the low-income group. The yoga intervention provided a two-level increment of the attribute “I’ve been feeling relaxed” for both groups. Even though the intervention promisingly improved the relaxation feeling for the high-income group towards the maximal level 5, fewer MWALY gains were generated when compared to the gains obtained by the low-income group. This example illustrates that increasing the score for this attribute from level 2 to level 4 was more valuable than increasing the score for this attribute from level 3 to level 5. The valuable magnitude would depend on the choice of tariff. It was more cost-effective to provide the yoga intervention for the low-income group than the high-income group, given that the costs per MWALY gain calculated by all tariff choices for the low-income group were lower. However, notably, the greatest discrepancy for costs per MWALY gain between the low-income group and high-income group was the one calculated using the DCE tariff based on mapping. The costs per MWALY gain calculated using this method for the low-income group was £97508.12 lower than that for the high-income group. The least discrepancy for costs per MWALY gain between the two groups was the one calculated using the C-TTO tariff, with only

£12423.49 lower for the low-income group. Implicitly, given that the value of improvement from level 2 to level 4 for the relaxation attribute was higher than that from level 3 to level 5, the improvement from level 2 to level 4 was the most valuable when the DCE tariff based on mapping was adopted.

7.4. Contributions of this research

As there is currently no preference-based tariff of a generic mental well-being instrument in the UK, the results of this thesis can inform future national valuation study of SWEMWBS. The derived valuation sets can be used to estimate MWALYs for the economic evaluation of mental well-being interventions, thus overcoming the sensitivity limitation of using QALYs to capture mental well-being benefits. Also, it was the first attempt to collect primary data in the qualitative and quantitative phases for testing the application of health state valuation techniques (C-TTO and DCE) in the valuation of mental well-being under a CAPI setting. The cognitive process and burden of completing the mental well-being valuation exercise and the modelling implications could be realised to understand the reliability of the valuation protocol in preference elicitation. Moreover, it was the first attempt to explore the application of the EQ-VT 2.1 protocol with adapted changes to the valuation of mental well-being. Lastly, due to the restriction of household mixing during the COVID-19 pandemic, Microsoft Teams was firstly adopted in the quantitative phase to explore its practicality in conducting virtual face-to-face interviews. The technical feasibility of enabling the remote-control function in Microsoft Teams for participants to complete the valuation tasks on their own was firstly explored in the quantitative phase. The use of remote-control function for self-completion purpose could reduce the possibility of participants tailoring answers to please the interviewer due to the interviewer clicking the answers on their behalf. This could be an advantage in ensuring the quality of valuation responses.

7.5. Limitations and directions for future research

As this thesis focuses on conducting piloting phases to check the robustness of a valuation protocol, the modelling results of this thesis were limited by the small sample size. Even though the coefficients generated by the modelling results in this thesis were in general coherent with reasonably small standard errors, future SWEMWBS valuation studies should aim to derive a national valuation set based on the preferences of the general

population from a larger sample size. With a sufficiently large sample size, enough statistical power can be maintained for applying more robust econometric techniques to model the C-TTO and DCE responses. More covariates (e.g. income level) and interaction parameters between the attribute levels and the demographic information can then be explored to analyse potential sub-group preferences. Second-order interaction terms between each of the attribute levels can also be explored to understand the influence of interaction effects between attribute levels on the preference elicitation results.

In addition to analysing different forms of heterogeneity on influencing preferences for mental well-being, future research can also continue to explore alternative ways of tackling potential non-intuitive or contradictory combinations of levels of SWEMWBS items. To minimise hypothetical bias, the most common way suggested in the previous studies for reducing the chance of encountering illogical states was to specify explicitly the exclusion of some potentially implausible states within the experimental design (Johnson *et al.*, 2013). This could ensure the absence of uncommon states within the generated choice sets. However, this method of manually tailoring the states included in the design could lose statistical efficiency in choice state generation. A trade-off between statistical efficiency and experimental realism was required. Another approach documented in the literature for dealing with the issue of illogical states was similar to the method used in my thesis, which was to opt for a choice set without highly uncommon states among many iterations (Yang *et al.*, 2019). For the selected choice set generated by the C-TTO and DCE experimental designs in the quantitative phase, the mental well-being states included within the choice set were double checked. It was confirmed that those highly uncommon states identified by the algorithm used in Appendix 17 were not appeared in the selected choice set. It was worth noting that this method of confirming the absence of highly uncommon reported states was successfully applied in the quantitative phase of this thesis, as the responses from the debriefing questions showed that nearly 80% of the participants somewhat agreed or strongly agreed with the absence of difficulty in imagining allocated mental well-being states. Apart from the experimental design strategies, I am aware that an alternative approach has been taken by some studies for selecting health states and modelling utility values to deal with the problem of presenting uncommon states to the participants (Young *et al.*, 2010). Conventional methods of modelling C-TTO or DCE utility values or selecting health states through experimental designs for a unidimensional measure were subject to limitations, because the assumption

of independent attributes was not satisfied. For example, SWEMWBS is a unidimensional measure. The seven items of SWEMWBS are correlated as they measure the same underlying construct of mental well-being. The resulting mental well-being states generated by the experimental designs could be difficult to imagine when uncommon combinations of levels of attributes were presented. This limitation could be accommodated by including all second-order interaction terms between attribute levels within the modelling specifications. However, the model specification would be complicated when too many independent variables were included. A potentially large sample size would be required to generate statistically significant results and ensure the predictive power of a model. To tackle these, instead of using conventional experimental designs (e.g. efficient design) for the selection of states for valuation, the concept of Rasch-based threshold analysis has been recently adopted by some preference studies for unidimensional measures with correlated items (e.g. unidimensional emotional component of the Clinical Outcomes in Routine Evaluation - 6D (CORE-6D)) to select a set of plausible states for valuations (Mavranezouli *et al.*, 2013). Rasch item threshold map was used to derive the most common response combinations of attribute levels across the position of a Rasch model logit scale. Regression models were then used to derive the relationship between mean TTO values and Rasch logit values. The valuation of the unidimensional mental health component for the ReQoL-UI adapted the idea of Rash-based health state selection and modelling approach to item response theory methods (Keetharuth *et al.*, 2021). A graded response model was used to estimating the probability of each possible combination of commonly encountered health states. Regression modelling was applied to estimate the relationship between TTO values and the item response theory based mental health score. Future research of SWEMWBS valuation studies can consider the use of these novel ways for the selection of mental well-being states for valuation and the modelling of SWEMWBS value set, given that SWEMWBS is a unidimensional measure with correlated items. The resulting value set could be compared with what have been presented in this thesis to investigate the quality of the generated value set. Future research could also explore the importance of eliminating implausible states in valuation exercise, as some evidence arguably suggested that the EQ-5D-5L valuation protocol or modelling results were not significantly affected by the inclusion of implausible states. Focusing on valuing common states could even mis predict the other states (Yang *et al.*, 2019).

Furthermore, future research can also focus on a comparison of costs per MWALY gain across datasets or the difference between costs per QALY gain and costs per MWALY gain within datasets. These could be used to explore policy implications of economic evaluations of mental well-being interventions. Specifically, a unit QALY and a unit MWALY are implying different things. For example, 1 QALY (e.g. living in full health for one year) is usually seen as ‘normal’ good health that might be experienced by anyone who does not have an illness. However, even when not suffering a specific impediment to well-being, a person might just be mentally well, but not necessarily experiencing full mental well-being. In other words, 1 MWALY (e.g. living in full mental well-being for one year) derived by the tariff sets reflects the highest possible well-being state, which is probably better than that a person typically experiences.

Also, it should be noted that the relationship between MWALYs and QALYs could be complementary, rather than substitutionary or completely separable. MWALYs could be more sensitive to capture the benefits of interventions related to mental well-being, whereas QALYs could be better in reflecting the benefits of physical health. For example, Powell *et al.* (2013) conducted a randomised controlled trial to investigate the effectiveness of web-based cognitive-behavioural tool in improving population mental well-being. WEMWBS was used as a primary outcome assessment. The results showed that there were statistically significant improvements of WEMWBS scores for the intervention arm at 6 weeks and 12 weeks. However, there were no statistically differences of the intervention arm at 6 weeks and 12 weeks for the EQ-5D scores, which are widely used in deriving the QoL adjustment weights to the QALYs. This patently revealed that the aspects of intervention benefits captured by MWALYs and QALYs in cost-utility analyses could be potentially different. MWALYs could be regarded as a generic outcome measure in the sense that improvements to any aspects of health result in improvements to mental health when the latter is measured with a measure like SWEMWBS that captures mental health improvements beyond the clinical range. Although SWEMWBS has been validated in mental illness populations, the value set generated by SWEMWBS in this thesis reflects the mental well-being preferences for the general population, not those of the condition-specific mental health patients, so the validity of the MWALY needs to be demonstrated in clinical populations. SWEMWBS has also not been well tested as a measure of other aspects of well-being (e.g. physical well-being). It may not be as sensitive as EQ-5D in capturing intervention benefits when evaluating an intervention

with a focus on improving physical aspects of health (e.g. a treatment to relieve pain of the patient). The role of MWALYs as a “common currency” in assessing “value for money” across interventions and conditions could be limited in this sense. Policy makers will need to decide whether costs per MWALY gain or costs per QALY gain would be more suitable in informing resource allocation decisions given the nature of interventions evaluated. Future research can continue to explore the correlation between QALY gains and MWALY gains associated with different natures of interventions. When judging whether an intervention is worth implementing, it could be too narrow a focus to investigate the underlying benefits solely in terms of either the physical health aspect or the mental health aspect. For example, even though a yoga intervention mainly aims to improve mental well-being of an individual (e.g. feeling more relaxed and thinking more clearly), we cannot completely neglect its physical health benefit as mental health status is shown to be correlated to morbidity and mortality (Barry *et al.*, 2009; Barry, 2009; Chida & Steptoe, 2008; Davidson, 2004; DiMatteo *et al.*, 2000; Friedli & Organization, 2009; Huppert & Baylis, 2004; Lyubomirsky *et al.*, 2005; Pressman & Cohen, 2005; Steptoe *et al.*, 2005). In this context, a high SWEMWBS score or a high MWALYs gain represents a high mental well-being, but it could be difficult to judge whether this refers to an overall “good” life. High mental well-being is just one of the elements for a “good” life. It will be worth investigating the causal relationship between improvement in physical health and improvement in mental well-being.

In terms of an enhanced generic instrument that can better reflect an individual overall well-being (but not restricted to mental well-being), there has been a recent development of the “EQ Health and Wellbeing Short (EQHWB-S)” (Mukuria *et al.*, 2021). The development of this instrument was motivated by the intention to capture benefits broader than health (i.e. extending beyond the QALY). The instrument has a balanced focus on both physical health and mental health assessment. The physical health construct is covered by the items related to difficulty doing day-to-day activities, difficulty getting around inside and outside, physical pain, and exhaustion. The mental health construct is covered by the items related to concentrating/thinking clearly, loneliness, sad/depression, anxiety, and the feeling of inability to cope with life. Future research is required to further validate the performance of this instrument in assessing well-being outcomes. It is noted that the items in EQHWB-S are predominantly negatively worded and do not cover some attributes (e.g. for mental well-being optimism and relationships with the others)

detectable by SWEMWBS. The possibility of developing an extended SWEMWBS instrument by including items related to physical well-being, and comparing this extended instrument with EQHWP-S could be an interesting future direction.

Moreover, due to limited time, the test-retest reliability or the completeness axiom of the C-TTO and DCE responses was not tested in this thesis. Future research can investigate the stability of the C-TTO and DCE responses in the long run by inviting same group of participants to answer the same set of C-TTO and DCE questions again after a certain period. Implications regarding the representativeness of the valuation answers overtime can then be inferred. This can also inform the time gap required to update the valuation set, as there can be dynamic changes in the notion and concept of mental well-being. The mental well-being preferences for the general population might vary from time to time. Also, both qualitative and quantitative phases in this thesis did not intentionally tailor the choice tasks to participants and the allocation of choice tasks was purely randomised with the aid of experimental design methods. In this context, other axioms of utility theory can be further investigated thoroughly in the future. For example, the monotonicity axiom can be tested during the completion of DCE tasks by tailoring a choice task with distinguishable difference in the level of mental well-being between alternatives. This can determine the number of participants who choose the dominated option (i.e. an alternative with all attribute levels relatively lower than the attribute levels of another alternative), which is a sign of irrationality. The continuity axiom can be investigated by observing whether participants always choose the alternative with the highest level of a specific attribute. A think-aloud interview can be used to help analyse their intuitions behind the selection behaviour with potential violation of utility axiom.

Apart from that, I am aware that the adoption of videoconferencing technology was used as an interview mode in a recent EQ-5D-5L preference elicitation study for Italy (Finch *et al.*, 2022). Future research of SWEMWBS valuation studies should continuously explore the technical feasibility of using virtual face-to-face interview as an online administration model. This thesis used Microsoft Teams for conducting the interviews in the quantitative phase. Even though most interviews were successfully administered under this online meeting platform, problems related to internet connection, software downloading, screen sharing, and remote-control function, etc. were observed. The effectiveness of using other alternative online meeting platforms such as Zoom and Skype

could be explored in the future. The possibility of combining virtual face-to-face interview and in-person face-to-face interview could be a cost-efficient way of collecting large-scale valuation responses.

Lastly, given that SWEMWBS is widely validated in the UK and some other countries, future research can continue to explore the psychometric robustness of SWEMWBS in different population sectors across the world. This can facilitate the derivation of country-specific SWEMWBS valuation set for health technology assessment.

7.6. Conclusion

This thesis documents two testing phases to investigate the validity of the C-TTO and DCE in the valuation of SWEMWBS. The results found that the application of these two health state valuation techniques is effective and suitable on reflecting individual preferences towards mental well-being. The preliminary versions of valuation set provide insights for the future derivation of a robust national valuation set in the UK.

Appendices

Appendix 1: Description of the MAU instruments

The QWB scale

The QWB scale, originally named the Health Status Index and the Index of Well-Being, was firstly developed in 1970 (Kaplan *et al.*, 1976). The construction of the scale was based on the General Health Policy Model (Fanshel & Bush, 1970; Kaplan & Anderson, 1996; Kaplan *et al.*, 1993), suggesting the measurement of both mortality and morbidity for the assessment of healthcare interventions by incorporating preferences for health states. The initial version of the QWB was composed of three aspects of functioning (mobility, physical activity and social activity) and it was administered by an interviewer. It was not commonly used by decision makers due to limitations of the scale, including the lack of mental health items, a huge cost of training for the interviewers, a long completion time, a potential recall bias when asking patients to recall symptoms and functions six days before the interview, etc. (Seiber *et al.*, 2008). These problems provided a driving force for the development of a modified version of the QWB.

- *QWB-SA*

It was developed with a view to addressing the limitations of the QWB. Instead of using interviewers as the administration mode, the QWB-SA is a self-completed and self-administered measure. The recall period was reduced to covering the past three days and the expected completion time was reduced to around 10 minutes. Also, the system checklist was expanded to include 19 chronic symptoms, 25 physical symptoms, 14 mental health symptoms and 17 items to cover the three dimensions of individual's mobility, physical activity, social and self-care activity, generating a total of 945 health states.

The visual analogue scale (VAS) was used as the valuation technique for the scoring of the QWB-SA, ranging from 0 (dead) to 100 (optimum health). The valuation weights were derived by the assumption of an additive model within the MAU Scaling method (Anderson & Zalinski, 1988), in which the weights obtained by the three dimensions were used to subtract the symptom weights. This application results in a lowest score of 0.09, which is lower than that of the QWB (0.33).

EQ-5D

The EQ-5D, which was first developed by the EuroQol Group in the 1990s, has long been a widely used health status instrument recommended and preferred by the NICE in England for eliciting preference (or utility) values for the health-related quality of life component of the QALY

(Devlin & Brooks, 2017; The National Institute for Health and Care Excellence, 2013). There are two main versions of the EQ-5D instrument: the EQ-5D-3L and the EQ-5D-5L.

- *EQ-5D-3L*

It is divided into two parts. The first part consists of a self-report health status classification system assessing respondents' health status on that day across five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has three levels indicating different levels of severity ((1) no problems; (2) some or moderate problems; and (3) severe or extreme problems), generating a total number of 243 possible health states. The second part of the measure is the EQ-VAS in which respondents are asked to indicate how good or bad their current health state is on a scale ranged from 0 (worst imaginable health state) to 100 (best imaginable health state). The EQ-VAS can be used to provide measurement implications of health states that are not captured by the health status classification system.

The valuation of the EQ-5D descriptive system is usually performed using a TTO method in which respondents are asked to indicate the amount of years of ill-health that they would be willing to trade to live in a state of full health. The scoring of the EQ-5D-3L is facilitated by a regression analysis in which the explanatory variables contain each dimension with different levels and an interactive "N3" term is incorporated to represent the occurrence of a level 3 response to at least one dimension. In the UK, the York A1 (Dolan) tariff is usually applied to each set of responses to generate an EQ-5D utility score (preference weight) for each health state. The utility values elicited range from -0.594 to 1, in which 1 represents the full health state and 0 represents the state of death (Dolan, 1997).

- *EQ-5D-5L*

Because of the substantial ceiling effect and concerns about the discriminatory power of the EQ-5D-3L, a new version of EQ-5D with 5 levels for each dimension has been developed. Similar to the EQ-5D-3L, the EQ-5D-5L contains the same five dimensions. However, two additional levels have been added to each of the five dimensions, generating a total of 3125 possible health states. The revised five levels of the EQ-5D-5L are "no problems", "slight problems", "moderate problems", "severe problems" and "extreme problems" or "unable". In addition, for the dimension of "mobility", the level-3 statement of the EQ-5D-3L descriptive system "I am confined to bed" is replaced by the level-5 statement of the EQ-5D-5L descriptive system "I am unable to walk about". Also, the word "performing" within the dimension of "usual activities" within the EQ-5D-3L descriptive system is replaced by "doing" within the EQ-5D-5L descriptive system. Other parts of the EQ-5D-5L descriptive system are the same as those of the EQ-5D-3L. Furthermore, the visual

analogue scale remains part of the EQ-5D-5L. However, in order to avoid confusion and facilitate score recording for the EQ-VAS, the EQ-5D-5L requires respondents to mark an “X” explicitly on the scale to indicate the position of the score and also write the score in the box provided.

The initial preference-based value set for the EQ-5D-5L has been reported by a study that invited participants to value a sample of the 3125 health states (Devlin *et al.*, 2018). A random sample of 996 adults were drawn from the general England population and each of them was asked to value 10 health states using a hybrid TTO and DCE methodology. Based on the most updated NICE position on the EQ-5D-5L, this published EQ-5D-5L valuation set in the England is still not recommended for reference-case analysis. The EQ-5D-3L valuation set should instead be used and the EQ-5D-5L utility values should be obtained by mapping onto EQ-5D-3L values, based on the mapping function developed by van Hout *et al.* (2012).

HUI

It is a generic preference-based measure that originated from the HUI1 classification system. It was originally mainly used to analyse the health outcomes of very low birthweight infants in neonatal intensive care (Boyle *et al.*, 1983; Cadman & Goldsmith, 1986; Torrance *et al.*, 1982). The HUI is currently divided into two main classification systems: HUI2 and HUI3. Both instruments are most widely used within North American populations within the context of HRQoL studies and economic evaluations (Horsman *et al.*, 2003; Richardson *et al.*, 2011b).

- *HUI2*

It was mainly designed for childhood studies and the questionnaire consists of seven attributes, namely “sensation”, “mobility”, “emotion”, “cognition”, “self-care”, “pain” and “fertility”. Each attribute is described by 3 to 5 levels, generating a total of 24,000 possible health states. It is noted that the attribute “fertility” is no longer assessed in the current HUI2 questionnaire.

A visual analogue technique and a SG approach is used to derive HUI2 preference scores using a multiplicative MAU function, with zero representing the health state of “death” and one representing the “perfect health” state. States worse than death are scored with negative values and the minimum score for the HUI2 is anchored at -0.03.

- *HUI3*

In contrast to the HUI2, the HUI3 is mainly designed for use by adult populations. There are eight attributes in total, namely: “vision”, “hearing”, “speech”, “ambulation”, “dexterity”, “emotion”, “cognition” and “pain”. Each attribute contains 5 to 6 levels of description, contributing to 972,000

possible health states in total. Compared to the HUI2, the HUI3 tries to improve the description details of the classification system. For example, the attribute “sensation” in the HUI2 is divided into two attributes “vision” and “hearing” in the new classification system. For the attribute of “emotion” within the HUI2, instead of incorporating 5 adjectives (fretful, angry, irritable, anxious, depressed) into a single level description, the HUI3 reduces the descriptive information to merely assessing the extent of happiness so as to avoid confusion and increase the preciseness of the description.

The valuation techniques for HUI3 health states apply the visual analogue scale and SG methods within single and MAU functions. Preference-based scores for the health state descriptions are elicited from the MAU functions in which the score represents a preference-based measure of health-related quality of life outcomes, with zero and one representing “dead” and “full health” respectively. Negative values are allowed for health states worse than death and the minimum score for the HUI3 is anchored at -0.36.

AQoL

It is the most common instrument in Australia for measuring the HRQoL of the general population (Anon, 2014). There are currently four versions of the AQoL instrument: AQoL-4D, AQoL-6D, AQoL-7D and AQoL-8D. The AQoL-4D was the first version of the AQoL in which only four dimensions were considered in the assessment of respondents’ health-related quality of life over the past week: “independent living”, “mental health”, “relationships” and “senses”. The AQoL-6D added two dimensions named “coping” and “pain” into the questionnaire, examining respondents’ feeling of the extent that they can solve the problems encountered in daily life and how often their feeling of pain or discomfort affects their usual activities. The AQoL-7D contains an additional dimension of “visual impairment” that assesses whether the vision of an individual is related to the ability of coping with the demands in life, ability to make friends and personal confidence to engage in daily activities, etc. Lastly, the AQoL-8D is the longest version of the AQoL and it expands the dimension space to incorporate “happiness” and “self worth”. It can theoretically generate 2.37×10^{23} health states for valuation. The scoring of the AQoL is simply derived by the summation of unweighted responses if it is used as a multi-attribute psychometric instrument. If it is used as a utility instrument, the responses are weighted to generate a preference-based utility score.

SF-6D

It is an instrument originating from a selection of items from the SF-36, which provides a longer version of the health status descriptive system. The SF-6D was developed by the University of

Sheffield and it is widely used within many research studies targeted at study populations within the UK and USA (Richardson *et al.*, 2011b). It has six dimensions in total: “physical functioning”, “role limitation”, “social functioning”, “pain”, “mental health” and “vitality”. Each dimension has between 4 and 6 levels, resulting in 18,000 possible health states. The preference-based scores for the health states have been valued by sample of the UK general public through the method of SG.

15D

It is a self-completed questionnaire designed for adults aged 16 or above and it is mostly used by studies that investigate the health-related quality of life in general and clinical samples of the Finnish population (Richardson *et al.*, 2011b). There are 15 dimensions in total that assess the respondent’s present health status: “mobility”, “vision”, “hearing”, “breathing”, “sleeping”, “eating”, “speech”, “excretion”, “usual activities”, “mental function”, “discomfort and symptoms”, “depression”, “distress”, “vitality” and “sexual activity”. Each dimension has between 4 and 5 levels, generating billions of health states potentially.

The valuation of health states is performed by VAS scaling and an additive model in the sense that the value of each response is multiplied by a utility weight to reflect the relative importance of the level within the corresponding dimension. An index score is then obtained by the summation of all valued states.

ReQoL

It is a generic mental health measure developed by the University of Sheffield and it is suitable for the completion of mental health service users aged 16 and above. There are two versions of ReQoL (ReQoL-10 and ReQoL-20) covering seven themes of recovery-focused quality of life outcomes: activity, hope, belonging and relationships, self-perception, well-being, autonomy, and physical health (Keetharuth *et al.*, 2018a). Both versions comprise of statements to assess the thoughts, feelings and activities of individuals with different mental health conditions over the last week.

- *ReQoL-10*

It contains 6 positively worded and 4 negatively worded mental health items, covering different constructs of hedonic (feelings) and eudaimonic (psychological functioning, relationships with the others and self-realisation) well-beings. The response categories for each item range from “none of the time” to “most or all of the time” within a five-point Likert scale. Also, there is one physical health item assessing the presence of physical problems. The description of five response categories for this item is phrased differently as “no problems”, “slight problems”, “moderate problems”, “severe problems”, and “very severe problems”.

This short version of questionnaire is handy for monitoring the recovery progress of service users continuously in routine clinical practice (Keetharuth *et al.*, 2017).

- *ReQoL-20*

On top of the ten mental health items and one physical health item included in the ReQoL-10, ReQoL-20 includes three extra positively worded and seven extra negatively worded mental health items. These constitute to the total of 20 mental health items and 1 physical health item in ReQoL-20 to explain the themes of recovery outcomes. This long version is mainly used in routine clinical practice and in research studies (Keetharuth *et al.*, 2017).

Both ReQoL-10 and ReQoL-20 are psychometrically similar (Keetharuth *et al.*, 2017; Keetharuth *et al.*, 2018b). The method of C-TTO with props was used to estimate the preference-based index for the ReQoL-Utility Index classification system, which consists of six mental health items and 1 physical health item selected from the seven themes. The derived utility values ranged from -0.195 to 1 (Keetharuth *et al.*, 2021).

Appendix 2: Description of the preference-based capability instruments

| |
|--|
| ICECAP-A |
| <p>It is used to assess the well-being of adults aged 18 or above based on Sen’s capability theory. The questions are divided into 5 attributes covering aspects of attachment (able to have love, friendship and support), stability (able to feel settled and secure), achievement (able to achieve and progress), enjoyment, and autonomy (able to be independent), with four levels for each attribute (Al-Janabi <i>et al.</i>, 2012). As this questionnaire aims to measure capability, the wording description for the four response categories contains the phrases “I am able to be” or “I can”. Best-worst scaling is the valuation method for scoring of the ICECAP-A in which 0 is equivalent to death with no capability and 1 indicates the best value with full capability on all aspects (Flynn <i>et al.</i>, 2007).</p> |
| ICECAP-O |
| <p>Instead of focusing on adult populations, the ICECAP-O aims to inform broader assessments of well-being other than health among older people who are aged 65 or above in the UK (Coast <i>et al.</i>, 2008a). Similar to the ICECAP-A, the questions of the ICECAP-O are classified into 5 attributes but there is a slightly different focus in terms of the coverage of well-being aspects: attachment (able to have love and friendship), security (able to think about the future without concern), role (able to do things that make you feel valued), enjoyment (able to have enjoyment and pleasure) and control (able to be independent). The four response categories for each attribute incorporate the phrases “I am able to be” or “I can”, as an emphasis on the assessment of ability and functionings. General population values for the capability states are derived by the best-worst scaling method and anchored on a scale ranging between 0 (no capability) and 1 (full capability).</p> |
| ASCOT |
| <p>It is used to investigate whether an individual’s needs and wants are fulfilled through the measurement of social-care related quality of life (Netten <i>et al.</i>, 2012). There are 8 domains in total within the questionnaire in which different kinds of basic and social needs are assessed: “accommodation cleanliness and comfort”, “safety”, “food and drink”, “personal care”, “control over daily life”, “social participation and involvement”, “dignity” and “occupation and employment”. Each domain category consists of four levels, with word phrasing of descriptions designed to measure capability. Respondents are assessed according to whether they are able to achieve or have their desired needs and wants fulfilled. This measure is valued by best-worst scaling or the use of the time trade-off method to elicit preference weights anchored on a scale ranging between 0 (being dead) and 1 (the ideal state).</p> |

Appendix 3: Description of the mental well-being instruments

| |
|--|
| <p>WEMWBS and SWEMWBS</p> <p>The WEMWBS was originally developed from the Affectometer 2, which is a 40-item mental health scale (20 positive and 20 negative) examining the positive mental health of an individual (Kammann & Flett, 1983). A total of 20 items are expressed as statements whilst the other 20 are represented as adjectives related to aspects of mental health such as “satisfied”, “optimistic” and “helpless”, etc. Although one study revealed that Affectometer 2 demonstrated an outstanding performance in terms of reliability, validity and acceptability among the UK population, the high internal consistency (Cronbach’s alpha = 0.944) suggested a potential reduction of the number of items included in this scale (Tennant <i>et al.</i>, 2007b). Because of this, improvements to the scale were made and it has been revised to a new scale called WEMWBS. It is a scale that consists of 14 positively worded statements assessing the positive mental health in terms of both hedonic and eudaimonic perspectives of well-being. Participants are given a five-point Likert scale ranging from “none of the time” to “all of the time” and they are asked to score each statement by referring to their personal experience and feelings over the past two weeks. The minimum and maximum scores for each statement are 1 and 5 respectively, with a total summing score ranging from 14 to 70. The higher the total score, the better the mental well-being of an individual, and vice versa.</p> <p>The Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS) is a 7-item scale, which is a shortened version of the WEMWBS. This shortened version has been created following a Rasch analysis of the WEMWBS that met the strict uni-dimensionality expectations of the Rasch model (Stewart-Brown <i>et al.</i>, 2009). It offers a more convenient alternative to WEMWBS in terms of brevity and conciseness.</p> |
| <p>WHO-5</p> <p>It is a short self-reported measure of mental well-being aimed at children and young people aged 9 years or above, covering the aspects of subjective well-being (Topp <i>et al.</i>, 2015). It was developed by the Psychiatric Research Unit, WHO Collaborating Centre in Mental Health in 1998. There are in total five positively worded statements assessing the feeling of respondents over the past two weeks, with response categories for each statement ranging from “at no time” (score of 0) to “all of the time (score of 5).</p> <p>The raw score is calculated by the summation of the scores of all five items, with 0 representing the worst mental well-being state and 25 representing the best mental well-being state. A percentage</p> |

score can also be obtained by multiplying the raw score by 4, yielding a possible range of scores from 0 to 100 in which a higher well-being is illustrated by a higher score, and vice versa.

MHC-SF

The MHC-SF, which was derived from the Mental Health Continuum-Long Form, is a 14-item and positively-worded questionnaire measuring the feeling and experience of respondents during the past month (Keyes, 2009). Specifically, there are three items assessing emotional (hedonic) well-being, six items assessing social (eudiamonic) well-being and five items assessing psychological (eudiamonic) well-being. Each item contains six response categories, ranging from “never (score of 0)” to “every day (score of 5)”.

The total summation score ranges from 0 to 70. A flourishing mental health is diagnosed when the response option of “every day” or “almost every day” is chosen for at least one of the 3 hedonic items and at least six of the 11 eudiamonic items. A languishing mental health is diagnosed when “never” or “once or twice” is chosen for at least one of the 3 hedonic items and at least six of the 11 eudiamonic items. A moderate mental health is diagnosed for those who fall into neither flourishing nor languishing mental health states.

SEHS

It is a 36-item survey developed mainly to assess the positive mental health of adolescents (Furlong, 2015; Furlong *et al.*, 2014). The survey is composed of four subscales: “Belief in self”, “belief in others”, “emotional competence” and “engaged living”, with nine items for each subscale. The response categories for 30 items are constructed as a 4-point scale ranged between “not at all true of me (score of 1)” and “very much true of me (score of 4)”, whereas a 5-point scale is designed for the remaining 6 items, ranged between “not at all (score of 1)” and “extremely (score of 5)”.

The total covitality score ranges from 36 to 150, and is calculated by summing scores from the four subscales. The higher the total score, the better the mental well-being of an individual, and vice versa. Different levels of covitality are represented by corresponding threshold scores: low (≤ 85), low average (86-106), high average (107-127) and high (≥ 128).

Appendix 4: Studies focusing solely on the validation of WEMWBS

| Population group | |
|---|---|
| Carers, adults, adolescents, parents, mental health patients | <p>Maheswaran <i>et al.</i> (2012) performed a secondary analysis of data obtained by the registered users of WEMWBS to statistically evaluate the responsiveness of WEMWBS in detecting a change in mental well-being conditions at both the individual and group levels in 12 intervention studies that had used the WEMWBS as an outcome measure. For the group level analysis, standardised response means were used for the evaluation of responsiveness through the investigation of probability of change statistic (\hat{P}). On the other hand, the standard error of measurement (SEM) was estimated to comprehend the responsiveness of WEMWBS at the individual level. Following investigation of the interventional studies, the general result revealed that the \hat{P} statistic for most studies was greater than 0.5 (lower limit of the 95% confidence interval (CI) for \hat{P}) while participants demonstrated an improvement in mental well-being at the 2.77 SEM threshold (lower limit of 95% CI > 5.0%) within all five studies being investigated at the individual level. These findings indicated a substantial responsiveness of WEMWBS at both the group level and individual level.</p> |
| English speaking Chinese or Pakistani | <p>Taggart <i>et al.</i> (2013) adopted a mixed methods study to evaluate the cross-cultural validation of the WEMWBS amongst English speaking people who identified themselves as Chinese or Pakistani in the UK. Quantitative data were collected in Birmingham and Coventry with the examination of psychometric properties in terms of normality, floor and ceiling effects, response rates, dimensionality, internal consistency and construct validity. Moreover, a qualitative evaluation was performed in Birmingham where mixed sex Chinese and single sex Pakistani focus groups were asked to determine their understanding and acceptability of the WEMWBS. Quantitative results basically demonstrated normally distributed scores for both the Chinese and Pakistan samples with no evidence of floor or ceiling effects. High response rates of 99% and 87% among the Chinese and the Pakistani samples, respectively, were achieved. A Cronbach's alpha of over 0.9 for both samples illustrated a good level of internal consistency and WEMWBS was consistent in assessing mental well-being. Exploratory</p> |

| | |
|---|--|
| | <p>factor analysis suggested a one-factor model for the explanation of variance. In terms of external consistency, WEMWBS showed a moderate and negative relationship with the GHQ-12 and a moderate and positive relationship with the WHO-5 within both the Chinese and Pakistani samples. These logical relationships between WEMWBS and other scales supported the construct validity of WEMWBS. Finally, qualitative results showed that the items within WEMWBS were generally user-friendly in terms of ease of completion and understanding. However, differences in cultural values and viewpoints among the Chinese and Pakistani groups caused occasional misinterpretations of the WEMWBS statements.</p> |
| <p>English and Scottish teenage students</p> | <p>Clarke <i>et al.</i> (2011) conducted a mixed methods assessment validating the WEMWBS among teenage students aged 13-16 years in six schools in England and Scotland. In addition to testing the psychometric properties of internal consistency and uni-dimensionality, this study examined the correlations of WEMWBS with comparator scales including the GHQ-12, WHO-5, Strengths and Difficulties Questionnaire (SDQ), MHC-SF and Kidscreen-27. Test-retest stability (or test-retest reliability) was also investigated by randomly selecting around 10% of participants to complete the WEMWBS a second time after one to two weeks. Eighty students also participated in a focus group to discuss their impressions of WEMWBS items. Similar to the findings of the other studies, WEMWBS was found to have strong internal consistency and a one factor model was confirmed by factor analysis. For the assessment of construct validity, moderate strength and statistically significant relationships were found between WEMWBS and the other comparator scales within the 95% confidence interval. The sign of the Spearman's rank correlation coefficients were shown to be positive between WEMWBS and MHC-SF (0.65), Kidscreen-27 (Physical Well-being: 0.43; Psychological Well-being: 0.59; Autonomy & Parent Relation: 0.46; Social Support & Peers: 0.38; School Environment: 0.51) and the WHO-5 (0.57), whilst it was negative between WEMWBS and SDQ (-0.44) and GHQ12 (-0.45) due to the latter's inversely related scoring criteria. The significant results indicated that WEMWBS measures a broad range of mental health, covering both hedonic and eudiamonic aspects of well-being. Regarding the reliability of the scoring results overtime, an intra-class correlation coefficient of 0.66 demonstrated an acceptable level</p> |

| | |
|--|--|
| | <p>of stable responses within a short period of time. Finally, when it came to the focus groups, participants generally found little to no difficulties in the completion of WEMWBS and most items were relevant to the measurement of mental well-being. However, the scale was subject to minor clarification as a few participants misunderstood the meaning of some WEMWBS statements such as the meanings of “feeling interested” for statement 4 (I’ve been feeling interested in other people) and “feeling close” for statement 9 (I’ve been feeling close to other people).</p> |
| <p>UK veterinary profession</p> | <p>Bartram <i>et al.</i> (2011) validated the WEMWBS across the UK veterinary profession by exploring the correlation of WEMWBS with the Hospital Anxiety and Depression Scale, Health and Safety Executive Management Standards Indicator Tool and questions regarding suicidal ideation. In addition to the absence of floor and ceiling effects of WEMWBS, the Pearson correlation coefficients revealed its statistically significant and strong negatively correlated relationship with the anxiety and depressive symptoms, supporting the divergent validity of the WEMWBS. Also, a mild-to-moderate and significant relationship of WEMWBS was found to be positively correlated with favourable psychosocial working conditions, supporting its convergent validity in the measurement of mental well-being. Moreover, the results obtained from the multiple logistic regression between WEMWBS and the other scales indicated that a 1 unit increase in WEMWBS score was associated with decreased odds of having suicidal thoughts, reporting anxiety and depressive symptoms.</p> |
| <p>Mexican youth</p> | <p>Hoffman <i>et al.</i> (2019) explored the dimensionality of the WEMWBS across youth from one of the schools in Mexico. Confirmatory factor analysis was used to investigate the fitness of a single-factor model to the response data of the scale. Although a statistically significant chi-square test indicated an evidence of data misfit, the other goodness-of-fit statistics (e.g. root mean square error of approximation (RMSEA) = 0.08; comparative fit index (CFI) = 0.94; Tucker-Lewis index (TLI) = 0.929) generally performed fine. Moreover, in terms of the standardised factor loading into the latent factor of mental well-being, the results are satisfactory in the sense that most items are above 0.4, with item 12 possessing the maximum loading (0.85) across all items. However, it is worth noting that the factor loadings of items 4 and</p> |

| | |
|---|--|
| | <p>11 were just around 0.35. Nevertheless, the precision of WEMWBS score within this group of Mexican youth was confirmed based on the high percentage of scoring variance attributable to the underlying factor. The general result regarding the uni-dimensionality of WEMWBS was acceptably reliable.</p> |
| <p>English speaking Pakistani healthcare professionals</p> | <p>Waqas <i>et al.</i> (2015) validated the psychometric properties of the WEMWBS among the English speaking Pakistani healthcare professionals, including physicians, surgeons, general practitioners and pharmacists, etc. The statistical results were favourable in general. Firstly, the practicality of the WEMWBS was supported by the achievement of 90% completion rate. Secondly, although a slightly skewed distribution was indicated by the item response frequencies, there was no evidence of floor or ceiling effects. Thirdly, based on the investigation of the Cattell's scree plot and eigenvalues, the principal components factor analysis revealed that the uni-dimensional structure of the WEMWBS could be assumed. Fourthly, the Cronbach's alpha of 0.89 indicated high level of internal consistency in the measurement of mental well-being, supplemented by the more than mild corrected item correlations for the 14 items. The content validity for the items was also verified by the moderate values of the item-total score correlations, which were acceptable. Fifthly, an excellent test-retest reliability was supported by an intra-class correlation coefficient of closer to 1, implying the lack of response fluctuation overtime. Finally, the readability of the WEMWBS was supported by the Flesch reading ease score of above 70 and the Flesch-Kincaid grade level score of around 4.5, revealing its ease of comprehension.</p> |

Appendix 5: Evidence focused solely on the validation of SWEMWBS

| Population group | |
|---|---|
| Cognitive hypnotherapy treatment for adults with common mental disorders | <p>Shah <i>et al.</i> (2021) investigated the performance of SWEMWBS in measuring the outcome of psychological treatment among patients with common mental disorders. Construct validity of SWEMWBS was assessed by its correlation with the Patient Health Questionnaire 9 (PHQ-9) and the General Anxiety Disorder 7 (GAD-7) scales. Internal consistency was assessed by the Cronbach’s Alpha. Time series analyses including within subject effects tests and within subject contrast tests were used to investigate the change or difference in the scores of three outcome measures for patients completing therapy sessions at different time points. The result showed that SWEMWBS was the only outcome measure with normally distributed score. High internal consistency was also demonstrated by the Cronbach’s Alpha of around 0.9 at different time points. Negative and significant correlations between SWEMWBS and PHQ-9 or GAD-7 were also discovered. The sign of coefficients was at the expected direction, as lower PHQ-9 and GAD-7 scores indicated less severe mental health problems. The change in SWEMWBS scores overtime was consistent at a linear trend. All these statistics properties supported the use of SWEMWBS in detecting the benefits of common mental disorders treatment.</p> |
| Welsh young people with different care status | <p>Anthony <i>et al.</i> (2021) explored the uni-dimensionality nature, measurement invariance properties and latent factor mean differences of SWEMWBS in the context of care status groups (foster, residential or kinship care placements). Secondary school students in years 7 to 11 with and without care of the local authority completed the SWEMWBS in the 2017 School Health Research Network Student Health and Wellbeing survey. The result supported the high internal consistencies of SWEMWBS, with the Cronbach’s Alpha above 0.8 across all care groups and not in care group. Categorical confirmatory factorial analysis confirmed the uni-dimensionality of SWEMWBS, as indicated by the values of all statistically significant standardised factor loadings greater than 0.5, and other model fit statistics (e.g. Comparative-of-Fit Index, Tucker–Lewis Index, Root Mean Square Error of Approximation). Moreover, the testing results of configural invariance, metric invariance, and scalar invariance suggested the invariant of SWEMWBS across all care groups. Furthermore, the latent mean</p> |

| | |
|--|---|
| | <p>comparisons revealed that young people in all care groups were associated with lower SWEMWBS scores, compared to young people without care.</p> |
| <p>Norwegian and Swedish hotel managers</p> | <p>Haver <i>et al.</i> (2015) validated the SWEMWBS across the Norwegian and Swedish hotel managers in a hotel chain. Data were obtained through an online completion of questionnaires, including the SWEMWBS. The general results supported the acceptable psychometric properties of the SWEMWBS. Firstly, it is acceptable that the skewness indices of both Norwegian and Swedish samples were -0.58 and -0.46 respectively, indicating an approximation symmetric distribution for the response data of SWEMWBS with slightly left skewed. Also, the internal consistencies for both samples were high, as indicated by the Cronbach's alpha of around 0.85. Besides, the uni-dimensionality of the SWEMWBS was supported by the high factor loadings ranging from 0.64 to 0.82 of all items within both samples, explaining more than 50% of the variance for one factor. The confirmatory factor analyses showed the moderate goodness of fit for the uni-dimensional structure of the SWEMWBS, supporting its factorial validity in loading to the latent variable of mental well-being. Moreover, in terms of criterion-related validity, the SWEMWBS was proved to be significantly mild-to-moderate and positively correlated with measures of mindfulness (the Mindful Attention Awareness Scale) and emotional intelligence (the Wong and Law Emotional Intelligence Scale). A relatively stronger and significantly positive correlation was obtained between SWEMWBS and the items of positive affect within the Positive and Negative Affect Schedule. Expectedly, a negative and significant correlation was shown between SWEMWBS and negative affect items. These results implied the positive framing measurement nature of the SWEMWBS. The discriminant validity was supported by no correlation between SWEMWBS and an unrelated measure, which was the number of rooms in the hotel managed by the managers in this article.</p> |
| <p>Deaf British Sign Language (BSL) users</p> | <p>This article examined the validity and reliability of the SWEMWBS among the Deaf BSL users in the UK after translating the English version of SWEMWBS into BSL (Rogers <i>et al.</i>, 2018). Participants from the Deaf community were asked to complete the SWEMWBS BSL, the CORE-OM BSL well-being subscale and the EQ-VAS from the EQ-5D BSL (Time 1), followed by the completion of the SWEMWBS BSL again after one week</p> |

| | |
|---|---|
| | <p>(Time 2). Good internal consistency for the SWEMWBS BSL was found, as indicated by the high Cronbach's alpha coefficients of 0.83 and 0.85 at Time 1 and Time 2 respectively. In terms of convergent validity, a statistically significant and negative Kendall's tau correlation coefficient was found between the SWEMWBS BSL and CORE-OM BSL whilst a statistically significant and positive correlation coefficient was found between the SWEMWBS BSL and EQ-5D VAS BSL. The results were valid and no contradictory relationships were found between these measures in terms of the sign of coefficient.</p> |
| <p>Singaporean mental health service users</p> | <p>Vaingankar <i>et al.</i> (2017) investigated the validity and reliability of SWEMWBS amongst 350 adult service users with schizophrenia, depression and anxiety spectrum disorders in Singapore. Psychometric properties including factorial validity, internal consistency, convergent and divergent validities were evaluated. The confirmatory factor analysis suggested the uni-dimensionality of SWEMWBS, as illustrated by the goodness-of-fit indices (CFI = 0.969; TLI = 0.954; RMSEA = 0.029). Furthermore, high internal consistency was indicated by a Cronbach's alpha coefficient greater than 0.7. Also, statistically significant moderate to high correlations above 0.6 were found between SWEMWBS and some convergent measures, including the Positive Mental Health and Satisfaction with Life Scale. It was evident that SWEMWBS measured aspects of mental wellbeing along with the other two scales. However, a weaker relationship was found between SWEMWBS and the Global Assessment of Functioning, as indicated by a coefficient of around 0.4. On the other hand, regarding divergent validity, SWEMWBS showed a moderate relationship with the Patient Health Questionnaire 8 and Generalised Anxiety Disorder (GAD-7), as estimated by a coefficient of around -0.5.</p> |
| <p>Mental health patients in Hong Kong</p> | <p>It examined the validity of SWEMWBS among patients with mental illness conditions including schizophrenia, personality disorder, bipolar affective disorder, depression and other problems in Hong Kong (Ng <i>et al.</i>, 2014). One hundred and twenty-six patients recruited in this study were asked to complete the Chinese version of SWEMWBS (C-SWEMWBS). Psychometric properties were investigated to discover whether the translated version of SWEMWBS is applicable for measuring the mental</p> |

| | |
|--|---|
| | <p>wellbeing of Chinese-speaking patients. The results displayed a normal distribution of C-SWEMWBS scores, demonstrating the discriminatory power of the instrument in distinguishing different levels of mental health conditions. A Cronbach's alpha of 0.89 was similar to that generated for the population sample in the UK. Five out of 7 items had a corrected item-total correlation above 0.7, supporting a strong internal consistency of C-SWEMWBS. In addition, among a subsample of 20 randomly selected individuals for the evaluation of test-retest reliability after two weeks of initial completion, the correlation between initial and retest mean scores was above 0.5. This demonstrated evidence of test-retest reliability over a short period of time. Furthermore, only a single component was found by the principal components factor analysis and it was consistent with the English version of SWEMWBS. Lastly, in terms of concurrent validity, there was a moderate relationship between the scores of C-SWEMWBS and WHO-5, as indicated by a correlation coefficient of 0.49. However, there was no statistically significant relationship between the scores of C-SWEMWBS and The Brief Psychiatric Rating Scale, implying that these two instruments may measure different attributes of mental health.</p> |
|--|---|

Appendix 6: Evidence focused on the validation of both WEMWBS and SWEMWBS

| Population group | |
|--|--|
| English mental health service users | <p>A myriad of studies have reported validation evidence for both the WEMWBS and SWEMEBS simultaneously. Bass <i>et al.</i> (2016) published a study that aimed to validate the WEMWBS within a population of secondary care mental health service users in the North East of England. Service users were asked to complete the WEMWBS and statistical analyses including Rasch analysis and a confirmatory factor analysis, which was used to test the reliability of the WEMWBS in this sample relative to the general population in the UK. Consistent with the previous finding of a high Cronbach’s alpha for the WEMWBS, this study also reported a Cronbach’s alpha of 0.9 for the SWEMWBS. This indicated that both versions performed well in terms of assessment of internal consistency. For the Rasch analysis, while most WEMWBS items were regarded as uni-dimensional in the measurement of mental well-being within the user sample, item 1 (“I’ve been feeling optimistic about the future”) and item 12 (“I’ve been feeling loved”) were misfits, as identified by outfit statistics of greater than 1.3. Furthermore, both the 7-item and 14-item versions demonstrated a CFI within the range of 0.95 and 0.97, with the RMSEA lying between 0.06 and 0.09. These results generally confirmed an acceptable fit to the one-factor model and it is suggested that WEMWBS is appropriate for measuring mental well-being amongst users of secondary care mental health services.</p> |
| English private households | <p>Ng Fat <i>et al.</i> (2017) used the data obtained from the Health Survey for England between 2010 and 2013 to evaluate the performance of SWEMWBS relative to the WEMWBS among a randomly selected representative sample of private households. SWEMWBS was assessed in terms of its correlation with the GHQ-12, EQ-VAS, happiness index in which participants were rated their level of happiness ranged from 0 (unhappy) to 10 (happy), self-rated health and limiting longstanding illness; this involved comparisons of correlation between WEMWBS and these instruments. The Cronbach’s alpha was also examined to assess the criterion validity of SWEMWBS. In terms of relative validity, the Bland-Altman method was adopted to explore the extent of agreement between the</p> |

| | |
|---------------------------------|---|
| | <p>scores of WEMWBS and SWEMWBS. Spearman correlation coefficients were estimated to analyse the association between these two scales. Also, weighted kappa statistics were used to explore a three-category version of SWEMWBS and WEMWBS. The general results indicated that the performance of SWEMWBS was similar to WEMWBS in the measurement of positive mental health. A moderate and statistically significant relationship was found between WEMWBS or SWEMWBS with the happiness index, GHQ-12 and EQ-VAS, as illustrated by spearman correlations ranging from 0.4 to 0.56 in absolute values. A relatively weaker correlation was found between WEMWBS or SWEMWBS and self-rated health and limiting longstanding illness, as shown by statistically significant coefficients of less than 0.36 in absolute values. Furthermore, the high Cronbach's alphas of 0.92 and 0.84 for WEMWBS and SWEMWBS, respectively, also revealed a substantial internal consistency for both scales. Additionally, the plot of the difference between WEMWBS and SWEMWBS scores against the mean of the two scores showed that there was no significant systematic difference between two scales as the line of equality was within the 95% CI. Regarding the result for relative validity, there was a statistically significant and positive relationship between SWEMWBS and WEMWBS within population subgroups stratified by sex, age, education level and income level. The correlation coefficients generally decreased slightly by 0.1 when estimating the relationship between SWEMWBS and the 7 items from WEMWBS that are not included in SWEMWBS. The interpretation of this change is that upward bias was resolved after the investigators excluded the repeated items between WEMWBS and SWEMWBS. Lastly, a general weighted kappa coefficient of above 0.8 indicated an almost perfect agreement between the categories of SWEMWBS and WEMWBS.</p> |
| <p>Danish population</p> | <p>Koushede <i>et al.</i> (2019) validated the psychometric properties of both Danish WEMWBS and SWEMWBS based on the data obtained from the Danish Mental Health and Well-being Survey 2016. The results generally supported the use of WEMWBS/SWEMWBS within the Danish population in the context of measuring mental well-being. Firstly, the content validity of both scales was supported by a normally distributed response scores and the absence of both floor and ceiling effects. Secondly, the result generated</p> |

from 11 cognitive interview revealed the general ease of completion and comprehension of scales, even though there was a few comments regarding the ambiguous and quirky of some item contexts, terminologies and response categories. Besides, the confirmatory factor analysis indicated a single-factor model for both WEMWBS and SWEMWBS, as reported by a number of goodness-of-fit statistics. For example, both CFI and TLI were greater than 0.95 and the RMSEA were around 0.06. Moreover, convergent validity was confirmed by statistically significant and positive correlations between the two scales and the WHO-5 (strongly correlated) and Self-rated health (moderately correlated) measured by the self-rating physical and mental health. Discriminant validity was confirmed by statistically significant and negative correlations between the two scales and the Patient Health Questionnaire for Depression and Anxiety 4, the Perceived Stress Scale and symptoms of discomfort and pain. Lastly, high internal consistencies were found for both WEMWBS and SWEMWBS, as indicated by the Cronbach's alpha of around 0.9.

Appendix 7: A review of the findings covering stage II to stage IV for the development process of a mental well-being preference-based instrument

| | Stewart-Brown <i>et al.</i> (2009) | Bartram <i>et al.</i> (2013) |
|--|---|---|
| Stage II: Eliminate and select the best items per dimension | <p>Results showed initial misfit to model expectations for items 8, 13 and 14 and the chi-square index for assessing the overall model fit was highly significant (p-value <0.001). Multidimensionality with gender bias and local dependency were also detected for some items in the WEMWBS, resulting in the total cancellation of 7 items in the scale. Although 2 of the remaining 7 items displayed differential item functioning for gender, the bias for these 2 items cancelled each other out. The resulting 7 items (Items 1, 2, 3, 6, 7, 9, 11) formed the establishment of SWEMWBS, which met the strict uni-dimensionality criteria of the Rasch model and demonstrated marginal improvement in fit.</p> <p>This was indicated by the goodness-of-fit statistics such as the accomplishment of nearer zero mean and standard deviation (SD) of 1 for the item-person fit statistics, when compared to the initial 14-item scale (14-item scale v.s. 7-item scale: item mean = 0.102 v.s. 0.065, item SD = 3.111 v.s. 1.341; person mean = -0.533 v.s. -0.475, person SD = 1.730 v.s. 1.222). The chi-square index for testing the item-trait interaction was statistically insignificant (p value >0.1), reflecting the fitness of the trait-groups to the Rasch model.</p> | <p>Consistent to the initial finding from the UK general population, an indication of initial misfit of the WEMWBS to model expectations was shown and the chi-square index was statistically significant (p-value <0.001). In terms of individual item fit, the fit residuals of items 5, 8, 10, 12, 13 and 14 highly deviated from an acceptable range of ± 2.5. Items 8 and 14 demonstrated the highest potential item redundancy, with the fit residuals of -10.59 and -9.61 respectively. Moreover, response to item 4 was shown to be biased by gender in the sense that females affirmed a higher response category than males for the given mental well-being level. It was deleted based on the uniform differential item functioning for gender. The cancellation of the 7 items resulted in the remaining items which are identical to the items of SWEMWBS. These remaining 7 items fitted to the model</p> |

| | | |
|--|---|---|
| | | <p>expectations with the evidence of uni-dimensionality, measurement invariance, minimum bias and minimum local dependency, etc.</p> <p>The item-person fit statistics were highly improved in the analysis of these 7 items, when compared to the initial 14-item scale (14-item scale v.s. 7-item scale: item mean = -0.574 v.s. -0.57, item SD = 5.341 v.s. 2.326; person mean = -0.540 v.s. -0.520, person SD = 1.644 v.s. 1.295). The item-trait interaction, as indicated by the chi-square index, was statistically insignificant (p value >0.1).</p> |
| Stage III: Explore item-level reduction | <p>The threshold map was used to examine the ordering of response categories. Generally, it revealed the absence of both floor and ceiling effects for the responses for all 14 items. The ordered thresholds demonstrated the change in mental well-being whenever there was a move from one level to another within each item. The number of thresholds (4) was one less than the number of levels (5) for each of the 14 items. There was no explicit evidence regarding the need to reduce the item levels.</p> | <p>The threshold ordering of the response options was investigated using the threshold map and the threshold probability curve to compare the logit ability rating of the item levels. Again, all the option thresholds were ordered, representing an increase in mental well-being when moving towards the next level.</p> |
| Stage IV: Validation: repeat stages I | <p>The robustness of the result obtained from the Rasch analysis was further tested in two alternative random samples of the full data set. The results fitted the expectations of the Rasch</p> | <p>The robustness of the results was verified by the cross-validation of four other data subsets. The results of the four</p> |

| | | |
|---|---|---|
| <p>to III on other data sets</p> | <p>model, demonstrating a mean of near 0 and a SD of close to 1 for the fitness of items and persons to the model (sample 1 v.s. sample 2: item mean = 0.126 v.s. 0.113, item SD = 0.681 v.s. 1.436; person mean = -0.472 v.s. -0.437, person SD = 1.223 v.s. 1.194). Also, the statistically insignificant chi-square interactions for both samples (p values >0.1) supported the absence of item-trait interactions.</p> | <p>data subsets generally fit the expectations of the Rasch model, as indicated by the item-fit and person-fit residuals (subset 1 v.s. subset 2 v.s. subset 3 v.s. subset 4: item mean = -0.065 v.s. 0.112 v.s. 0.511 v.s. -0.054, item SD = 1.923 v.s. 1.473 v.s. 2.852 v.s. 2.777; person mean = -0.476 v.s. -0.389 v.s. -0.287 v.s. -0.516, person SD = 1.222 v.s. 1.933 v.s. 1.342 v.s. 1.459). Although one data subset showed a statistically significant chi-square index at the 5% significant level (p value = 0.041), statistically insignificant results for the chi-square indices were found for all four data subsets when restricting the rejection rule at the 1% significant level.</p> |
|---|---|---|

Appendix 8: Descriptive system of the SWEMWBS

Below are some statements about feelings and thoughts.

Please tick the box that best describes your experience of each over the last 2 weeks

| STATEMENTS | None of the time | Rarely | Some of the time | Often | All of the time |
|--|------------------|--------|------------------|-------|-----------------|
| I've been feeling optimistic about the future | 1 | 2 | 3 | 4 | 5 |
| I've been feeling useful | 1 | 2 | 3 | 4 | 5 |
| I've been feeling relaxed | 1 | 2 | 3 | 4 | 5 |
| I've been dealing with problems well | 1 | 2 | 3 | 4 | 5 |
| I've been thinking clearly | 1 | 2 | 3 | 4 | 5 |
| I've been feeling close to other people | 1 | 2 | 3 | 4 | 5 |
| I've been able to make up my own mind about things | 1 | 2 | 3 | 4 | 5 |

"Short Warwick Edinburgh Mental Well-Being Scale (SWEMWBS)

© NHS Health Scotland, University of Warwick and University of Edinburgh, 2008, all rights reserved."

| Name of items | Response categories |
|---|---|
| I've been feeling optimistic about the future | <ol style="list-style-type: none"> 1. None of the time feeling optimistic about the future 2. Rarely feeling optimistic about the future 3. Some of the time feeling optimistic about the future 4. Often feeling optimistic about the future 5. All of the time feeling optimistic about the future |
| I've been feeling useful | <ol style="list-style-type: none"> 1. None of the time feeling useful 2. Rarely feeling useful 3. Some of the time feeling useful 4. Often feeling useful 5. All of the time feeling useful |
| I've been feeling relaxed | <ol style="list-style-type: none"> 1. None of the time feeling relaxed 2. Rarely feeling relaxed 3. Some of the time feeling relaxed |

| | |
|--|--|
| | <ul style="list-style-type: none"> 4. Often feeling relaxed 5. All of the time feeling relaxed |
| I've been dealing with problems well | <ul style="list-style-type: none"> 1. None of the time dealing with problems well 2. Rarely dealing with problems well 3. Some of the time dealing with problems well 4. Often dealing with problems well 5. All of the time dealing with problems well |
| I've been thinking clearly | <ul style="list-style-type: none"> 1. None of the time thinking clearly 2. Rarely thinking clearly 3. Some of the time thinking clearly 4. Often thinking clearly 5. All of the time thinking clearly |
| I've been feeling close to other people | <ul style="list-style-type: none"> 1. None of the time feeling close to other people 2. Rarely feeling close to other people 3. Some of the time feeling close to other people 4. Often feeling close to other people 5. All of the time feeling close to other people |
| I've been able to make up my own mind about things | <ul style="list-style-type: none"> 1. None of the time able to make up my own mind about things 2. Rarely able to make up my own mind about things 3. Some of the time able to make up my own mind about things 4. Often able to make up my own mind about things 5. All of the time able to make up my own mind about things |

Appendix 9: A review of the direct valuation techniques

| Direct valuation technique | Description |
|---|--|
| VAS (also referred as the category rating scale or the rating scale) | <p>A line with well-defined end points representing the best imaginable (or the most preferred) health states and the worst imaginable (or the least preferred) health states at either end. Respondents are asked to indicate the position of the scale reflective of their own current health condition. One of the most common VAS scales was produced by the EuroQol research group in which a rating thermometer ranged from 0 to 100 is used to assess the health state of respondents on a particular day. However, there are many variants of the VAS as the length of the line could vary and it could be presented as either a vertical or a horizontal line. It can also be shown with the absence of interval scores marked explicitly. Nevertheless, the VAS is regarded as an interval scale in the sense that the difference between 20 and 25, using the example of the EQ-VAS, is considered the same as the difference between 85 and 90 (Kaplan <i>et al.</i>, 1979).</p> |
| ME | <p>An alternative valuation technique to the VAS. Rather than asking respondents to make a choice between two options, they are asked to compare health states in terms of ratio scaling. The ratio of undesirability is obtained by indicating how many times (x) is health state A worse than health state B, with a view to inferring the number of times (x) the disutility of health state A is as great as that of health state B (Torrance, 1986).</p> |
| SG | <p>A technique based on the theory of expected utility in which respondents are asked to make a choice under uncertainty, with the fulfilment of a set of utility axioms regarding individual preferences (Morgenstern & Von Neumann, 1953). Utility values are estimated for possible health outcomes and individuals decide the most desirable option that maximises their expected utility. In the context of the SG in health state valuation, according to whether the health state is regarded as better or worse than death, respondents are given two alternatives: a specific certain outcome is represented by one alternative whereas a risky gamble with two possible outcomes is described by another alternative.</p> <p>For a health state h_i which is worse than full health but preferred to death, the two alternatives are presented to the respondents as follows:</p> |

| | |
|------------|--|
| | <ul style="list-style-type: none"> ➤ Alternative 1 (Uncertain option) – There is a treatment with probability p that the respondent will return to the full health and live for an additional t years before death, and a probability $1-p$ of immediate death. ➤ Alternative 2 (Certain option) – The respondent will die after living in health state h_i for t years. <p>The probability p is varied until individual respondent indicates indifference between the two alternatives. The probability p at the indifference point is equivalent to the utility value for health state h_i.</p> <p>For a health state h_i which is considered worse than dead, the two alternatives are presented to the respondents as follows:</p> <ul style="list-style-type: none"> ➤ Alternative 1 (Uncertain option) – There is a treatment with probability p that the respondent will return to the full health and live for an additional t years before death, and a probability $1-p$ of staying in a health state h_i for t years before death. ➤ Alternative 2 (Certain option) – The respondent will die immediately. <p>The probability p is varied until the individual respondent indicates indifference between the two alternatives. The utility value of the health state h_i is equivalent to the formula $h_i = -p / (1 - p)$, where the utility value is bounded between $-\infty$ and $+1$.</p> |
| TTO | <p>It was developed by Torrance (1976) as an alternative to the SG. Instead of presenting an alternative with uncertainty to respondents, the TTO presents two certain outcomes to respondents. Because of this, it is regarded as a less complicated valuation technique relative to the SG as the difficulty of explaining probability theory to respondents is avoided and the confusion generated by theoretical understandings can be minimised. Within the TTO framework, respondents are required to choose between two alternatives based on the concept of opportunity cost, in which the respondents can obtain a better health status only with the cost of having a shorter life span. Similar to the SG valuation exercise, there are variations for the two alternatives based on whether the health state is regarded as either better or worse than death.</p> |

| | |
|--|---|
| | <p>For a health state h_i which is worse than full health but preferred to death, the two certain alternatives are presented to the respondents as follows:</p> <ul style="list-style-type: none"> ➤ Alternative 1 – The respondent lives in health state h_i for a period of t before death. ➤ Alternative 2 – The respondent lives in full health for a period of x, where $x < t$. <p>The time period x is varied until the individual respondent indicates indifference between the two alternatives. The valuation score given to this indifference point is calculated as $h_i = x/t$.</p> <p>For a health state h_i that is considered worse than death, the two certain alternatives are presented to respondents as follows:</p> <ul style="list-style-type: none"> ➤ Alternative 1 – The respondent dies immediately. ➤ Alternative 2 – The respondent lives in health state h_i for a period of y, followed by a period of x in full health, where $x + y = t$. <p>The time period x is varied until the individual respondent indicates indifference between the two alternatives. The valuation score given to this indifference point is calculated as $h_i = -x/(t-x)$, where the valuation score is bounded between $-\infty$ and $+1$.</p> <p>As noted in both the SG and TTO valuation techniques, there is no lower bound for the negative valuation score generated by the health state considered worse than death. The imbalance weighting between the health state considered better than death and the health state considered worse than death is problematic due to the presence of a negative bias towards the negative value. Recently, two alternative versions of the conventional TTO have been developed to tackle this valuation limitation (Robinson & Spencer, 2006).</p> |
| <p><i>Lead-time time trade-off (Lead-time TTO)</i></p> | <p>This version of the TTO incorporates a period of full health at the beginning of the conventional TTO task (lead time), no matter whether the health state is regarded as better or worse than death (Devlin <i>et al.</i>, 2011). Theoretically, for a health state h_i that is worse than full health but preferred to death, the two certain alternatives are presented to the respondents as follows:</p> |

| | |
|--|--|
| | <ul style="list-style-type: none"> ➤ Alternative 1 – The respondent lives in full health for a period of g years before death. ➤ Alternative 2 – The respondent lives in full health for a period of f years, followed by the health state h_i for a period of $t-f$ before death, where $g > f$. <p>For a health state h_i that is considered worse than death, the two certain alternatives are presented to the respondents as follows:</p> <ul style="list-style-type: none"> ➤ Alternative 1 – The respondent lives in full health for a period of g years before death. ➤ Alternative 2 – The respondent lives in full health for a period of f years, followed by the health state h_i for a period of $t-f$ before death, where $g < f$. <p>In both cases, the time period g is varied until the individual respondent indicates indifference between the two alternatives. The valuation score given to this indifference point is calculated as $h_i = (g-f) / (t-f)$, where the valuation score is bounded between -1 and $+1$ when the ratio of lead time to health state is equal to $1:1$.</p> |
| <p><i>Lag-time time trade-off (Lag-time TTO)</i></p> | <p>The reversed version of the lead-time TTO in which a period of full health is added after instead of before the health state h_i (Augustovski <i>et al.</i>, 2013). In other words, for a health state h_i which is worse than full health but preferred to death, the two certain alternatives are presented to the respondents as follows:</p> <ul style="list-style-type: none"> ➤ Alternative 1 – The respondent lives in full health for a period of g years before death. ➤ Alternative 2 – The respondent lives in the health state h_i for a period of $(t-f)$ years, followed by full health for a period of f before death, where $g > f$. <p>For a health state h_i that is considered worse than death, the two certain alternatives are presented to the respondents as follows:</p> <ul style="list-style-type: none"> ➤ Alternative 1 – The respondent lives in full health for a period of g years before death. ➤ Alternative 2 – The respondent lives in the health state h_i for a period of $(t-f)$ years, followed by full health for a period of f before death, where $g < f$. <p>The valuation criteria for reaching an indifference point and valuation formula are the same as that of the lead-time TTO. The valuation score is bounded between -1 and $+1$ when the ratio of lag time to health state is equal to $1:1$. It is noted that</p> |

| | |
|---|---|
| | <p>alternative 1 in the above cases of the lag-time TTO are identical to those of the lead-time TTO, regardless of whether the health state h_i is considered better than or worse than death.</p> |
| <p>PTO (originally named as the equivalence technique)</p> | <p>A choice-based technique used in the context of social decision making in which respondents are asked to choose between alternatives that involve other people, rather than deciding about their own or hypothetical health state. Due to this feature, a broader aspect of social value or well-being value for health states can be estimated. A sample question framing is structured as follows and the alternatives are usually provided under the assumption of equal cost or a fixed budget:</p> <p><i>“If there are x people in adverse health situation A and y people in adverse health situation B, and if you can only help (cure) one group, which group would you choose?” (Prades, 1997; Richardson, 1994)</i></p> <p>The number y is varied until the respondent reaches an indifference point between two groups in terms of needing help. The valuation result can be expressed within an undesirability scale, given that the undesirability of health situation B is x/y times as large as that of health situation A.</p> |

Appendix 10: A review of the indirect valuation techniques

| Indirect valuation technique | Description |
|-----------------------------------|---|
| Mapping (or cross-walking) | <p>A statistical technique used to derive utility values for a non-preference-based instrument from a generic preference-based instrument through the estimation of the statistical association between these two instruments using different forms of regression techniques (Brazier <i>et al.</i>, 2010). The exchange rates between them can then be obtained. However, the validity or the power of the mapping result depends largely on the nature of the mapped instruments. The strength of mapping can be maximised only when there is a sufficiently high degree of overlap between the two instruments in terms of attributes or dimensions as the constructs covered by one instrument might not be captured by another instrument. For instance, the crosswalk value estimated from mapping between a non-preference-based condition-specific instrument and a preference-based generic instrument may be limited by the level of item correlation.</p> |
| DCEs | <p>A stated preference technique in which a preference-based value is elicited through choosing between pairs of profiles or multiple options (Lancsar & Louviere, 2008; Ryan, 2004). Each profile consists of a combination of levels of different attributes and respondents are asked to indicate the most preferred option across different profiles presented. The theoretical grounding of the utility value derived based on the choice made by the respondents lies in random utility theory, which states that the utility sum of individual i conditional on choice j is formulated as follows (Hanemann, 1984; Lancaster, 1966; McFadden, 1973):</p> $U_{ij} = V_{ij} + \varepsilon_{ij},$ <p>where $V_{ij} = X'_{ij}\beta + Z'_i\gamma$</p> <p>$V_{ij}$ is the systematic or explained part composed of a vector of attributes of good j observed by individual i (X'_{ij}) and a vector of characteristics of individual i (Z'_i). β and γ are the vectors of estimated coefficients. ε_{ij} is a random or unexplained component caused by some unobserved characteristics such as differences in personal taste and attitude. This form of utility measurement is based on the idea</p> |

| | |
|-------------------|---|
| | <p>that the satisfaction gained from consumption does not only depend on the amount of goods and services possessed, but also the weighting of different attributes attached to a particular good. However, the resulting values obtained by the DCE are expressed on a latent utility scale and further methods such as hybrid methods and mapping DCE values onto TTO utilities are required to convert latent values onto the full health-dead scale (Rowen <i>et al.</i>, 2015).</p> |
| <p>BWS</p> | <p>An ordinal data collection technique which was first introduced to the health economics research area in the early 21st century (Flynn <i>et al.</i>, 2007; McIntosh & Louviere, 2002). Fundamentally, a profile case with a list of at least three aspects of health states from different levels of the measurement instrument is presented to respondents. They are asked to imagine living in the presented health states and indicate their choices of the best state and the worst state. Specifically, there are currently three types of BWS. The first one is the BWS case 1 (or “object” case). A list of attributes are presented to respondents and they are asked to indicate their most and least preferred attributes. The second type is called the BWS case 2 (or “profile” case), in which a list of attributes’ levels are given for respondents to indicate the most and least preferred levels. The third type is called the BWS case 3 (or “multi-profile” case), in which different options of alternatives with various attributes and the corresponding levels are presented to the respondents. They are asked to indicate the most and least preferred profiles of alternatives. The utility value derived from this choice elicitation task is also based on the theory of random utility. Respondents are assumed to maximise the difference in utility between the two attribute levels in every pair of attributes (Flynn <i>et al.</i>, 2007; Marley & Louviere, 2005).</p> |

Appendix 11: An example of the advertisement layout in the qualitative phase

Study title: Valuation of mental well-being as measured by the Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS): A think-aloud interview study

Are you interested in a mental well-being study? Do you want to contribute your opinions towards the development of valuation methods for mental well-being? Here is the chance!

- **Study aim:** To test approaches developed by economists for valuing health states and see how well they apply to valuing mental well-being states. Your participation will help us understand preferences for different approaches.

What will be done during the interview?

- To make trade-off between choices of imaginable life with different durations of mental well-being status.
- To look at pairs of mental well-being profiles and choose the one you prefer.
- To give us your views about the feelings and thoughts of completing the valuation exercises.



Name of interviewer: Hei Hang Edmund Yiu, Ph.D. student at Warwick Medical School, University of Warwick

Interview duration: ~60 minutes

Venue: A central local area, a workplace or your home in Warwickshire & West Midlands.

Contact email: H.Yiu@warwick.ac.uk

Mobile number: 07596212441

Please get in touch by email or phone if you would like to participate. I look forward to hearing from you.

Appendix 12: Syntax for the DCE experimental design in Ngene

```
Design
;alts = alt1, alt2
;rows = 32
;eff = (mnl,d)
;block = 4
;model:

U(alt1) = b0[0] + b1.dummy[0|0|0|0]*item1[1,2,3,4,5] +
b2.dummy[0|0|0|0]*item2[1,2,3,4,5] + b3.dummy[0|0|0|0]*item3[1,2,3,4,5]
+ b4.dummy[0|0|0|0]*item4[1,2,3,4,5] +
b5.dummy[0|0|0|0]*item5[1,2,3,4,5] + b6.dummy[0|0|0|0]*item6[1,2,3,4,5]
+ b7.dummy[0|0|0|0]*item7[1,2,3,4,5] /

U(alt2) = b1.dummy*item1 + b2.dummy*item2 + b3.dummy*item3 +
b4.dummy*item4 + b5.dummy*item5 + b6.dummy*item6 + b7.dummy*item7

$
```

Appendix 13: The 32 pairs of MWB states included in the DCE valuation tasks

| Block | Alternative 1 | | | | | | | Alternative 2 | | | | | | |
|--------------|----------------------|---|---|---|---|---|---|----------------------|---|---|---|---|---|---|
| 1 | 4 | 1 | 3 | 5 | 5 | 3 | 1 | 5 | 2 | 4 | 1 | 1 | 4 | 2 |
| 1 | 4 | 4 | 4 | 2 | 1 | 5 | 4 | 3 | 2 | 3 | 5 | 4 | 1 | 2 |
| 1 | 5 | 2 | 3 | 2 | 2 | 5 | 4 | 4 | 4 | 1 | 4 | 1 | 4 | 3 |
| 1 | 5 | 3 | 1 | 4 | 3 | 3 | 3 | 1 | 1 | 2 | 3 | 2 | 5 | 5 |
| 1 | 4 | 5 | 4 | 1 | 5 | 3 | 2 | 5 | 1 | 1 | 5 | 4 | 2 | 4 |
| 1 | 3 | 2 | 2 | 4 | 1 | 3 | 4 | 1 | 4 | 5 | 1 | 5 | 2 | 1 |
| 1 | 2 | 2 | 4 | 2 | 4 | 2 | 1 | 1 | 4 | 3 | 4 | 1 | 5 | 2 |
| 1 | 1 | 2 | 1 | 3 | 1 | 3 | 1 | 4 | 5 | 2 | 2 | 4 | 1 | 4 |
| 2 | 1 | 1 | 5 | 4 | 2 | 2 | 1 | 3 | 2 | 1 | 2 | 5 | 5 | 5 |
| 2 | 2 | 5 | 2 | 1 | 3 | 4 | 4 | 4 | 3 | 5 | 2 | 1 | 2 | 3 |
| 2 | 2 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | 5 | 1 | 2 | 2 | 3 | 5 |
| 2 | 4 | 4 | 1 | 3 | 2 | 1 | 1 | 1 | 5 | 3 | 2 | 3 | 2 | 3 |
| 2 | 2 | 4 | 2 | 2 | 3 | 1 | 3 | 3 | 3 | 4 | 4 | 5 | 2 | 2 |
| 2 | 3 | 1 | 4 | 3 | 4 | 5 | 3 | 2 | 3 | 2 | 4 | 2 | 2 | 5 |
| 2 | 1 | 5 | 1 | 3 | 5 | 1 | 4 | 2 | 1 | 4 | 1 | 4 | 3 | 5 |
| 2 | 3 | 3 | 5 | 5 | 2 | 3 | 2 | 4 | 2 | 3 | 4 | 4 | 4 | 1 |
| 3 | 5 | 5 | 1 | 1 | 4 | 2 | 5 | 3 | 2 | 5 | 2 | 3 | 4 | 1 |
| 3 | 4 | 2 | 5 | 5 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 2 | 4 | 3 |
| 3 | 1 | 1 | 1 | 2 | 5 | 2 | 2 | 5 | 5 | 5 | 3 | 1 | 1 | 1 |
| 3 | 1 | 3 | 2 | 5 | 1 | 4 | 5 | 2 | 4 | 1 | 4 | 2 | 5 | 2 |
| 3 | 5 | 3 | 5 | 2 | 4 | 4 | 2 | 3 | 4 | 4 | 5 | 3 | 2 | 4 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 5 | 3 | 3 | 2 | 2 | 5 | 5 | 1 | 2 | 1 | 5 | 1 | 1 |
| 3 | 3 | 2 | 5 | 1 | 3 | 5 | 5 | 2 | 1 | 4 | 2 | 1 | 1 | 2 |
| 3 | 2 | 4 | 3 | 4 | 5 | 1 | 5 | 4 | 5 | 2 | 5 | 3 | 5 | 2 |
| 4 | 3 | 4 | 2 | 2 | 4 | 2 | 2 | 2 | 3 | 4 | 5 | 2 | 5 | 3 |
| 4 | 1 | 3 | 4 | 4 | 3 | 1 | 5 | 2 | 5 | 3 | 1 | 1 | 3 | 1 |
| 4 | 2 | 5 | 2 | 4 | 1 | 5 | 1 | 4 | 2 | 3 | 1 | 2 | 3 | 4 |
| 4 | 1 | 2 | 2 | 1 | 4 | 5 | 3 | 5 | 1 | 1 | 3 | 3 | 4 | 5 |
| 4 | 3 | 1 | 1 | 1 | 1 | 4 | 4 | 5 | 2 | 2 | 3 | 5 | 3 | 3 |
| 4 | 4 | 1 | 3 | 1 | 2 | 1 | 3 | 1 | 4 | 5 | 5 | 4 | 3 | 4 |
| 4 | 5 | 4 | 4 | 5 | 2 | 4 | 1 | 2 | 1 | 5 | 3 | 5 | 1 | 4 |
| 4 | 5 | 1 | 5 | 5 | 5 | 4 | 3 | 1 | 3 | 1 | 1 | 3 | 1 | 1 |

Appendix 14: Code for the C-TTO experimental design in R

```
rm(list=ls())

# load files with states format should be: one column per dimension;
header row with dimension name; each row includes one state

# TTO_fixed contains the 7 mildest states and the worst state

# TTO_all contains all states except the worst and the 7 mildest

TTO_fixed <- read.csv("E:/Warwick/PhD/Preference
tariff/Thesis/Experimental design/Composite TTO/TTO_fixed_states.csv")

TTO_all <- read.csv("E:/Warwick/PhD/Preference
tariff/Thesis/Experimental design/Composite TTO/TTO_all_states.csv")

num_fix <- 8
num_var <- 42
num_all <- 78117

lvldistmat <- array(0,dim=c(10,7))

# set an initial big number, for comparison purpose. If the generated
level balance check is smaller than this big number, will go for the
generated level balance check, then continue to the second loop.

lvlbalcheck <- 10000000000

lvldistmatbest <- array(0,dim=c(10,7))

TTOrandbest <- rep(NA,42)

# number of iterations is still 10,000.
for (m in 1:10000){

  EQ <- array(0,dim=c(num_fix+num_var,7))
  EQlvlmat <- array(0,dim=c(5,7))

  # select random subset of "num_var" states from TTO_all which
includes "num_all" states in total

  TTO_rand <- TTO_all[sample(num_all,num_var,F),]
```

```

dim(TTO_rand)

# combine the sampled and fixed states into a single design
TTO <- rbind(TTO_fixed,TTO_rand)

# check design for level balance

for(i in 1:7){
  EQ[,i] <- TTO[,i]
}

for (j in 1:7) {
  for (k in 1:5) {
    EQlvlmat[k,j]<-sum(EQ[,j]==k)
  }
}

for (j in 1:7) {

  lvlldistmat[1,j]<-(EQlvlmat[1,j]-EQlvlmat[2,j])^2
  lvlldistmat[2,j]<-(EQlvlmat[1,j]-EQlvlmat[3,j])^2
  lvlldistmat[3,j]<-(EQlvlmat[1,j]-EQlvlmat[4,j])^2
  lvlldistmat[4,j]<-(EQlvlmat[1,j]-EQlvlmat[5,j])^2

  lvlldistmat[5,j]<-(EQlvlmat[2,j]-EQlvlmat[3,j])^2
  lvlldistmat[6,j]<-(EQlvlmat[2,j]-EQlvlmat[4,j])^2
  lvlldistmat[7,j]<-(EQlvlmat[2,j]-EQlvlmat[5,j])^2

  lvlldistmat[8,j]<-(EQlvlmat[3,j]-EQlvlmat[4,j])^2
  lvlldistmat[9,j]<-(EQlvlmat[3,j]-EQlvlmat[5,j])^2
}

```

```

    lvldistmat[10,j]<-(EQlvlmat[4,j]-EQlvlmat[5,j])^2
  }

  lvlbalcheck2 <- sqrt(sum(lvldistmat[,]))

  if (lvlbalcheck2<lvlbalcheck){
    lvlbalcheck <- lvlbalcheck2
    lvldistmatbest <- lvldistmat
    TTOrandbest <- TTO_rand
  }

  ###lvlbalcheck
}

head(lvldistmatbest)
lvlbalcheck

##### Blocking

library("AlgDesign")

des <- TTOrandbest

## number of pairs per respondent
npairs <- 6

## number of blocks
nblocks <- 7

desBlock<-optBlock(~.,des,c(rep(npairs,nblocks)))
print(desBlock$Blocks)

```

```
TTO_blocked<-desBlock$Blocks

TTO_blocked
rows<-sample.int(7,7,replace=FALSE)
for ( i in 1: 7){

  mx <- rbind(mx,TTO_fixed[rows[i],])

}
mx
mx <- TTO_fixed[2:8,]
mx2 <- mx
for ( i in 1: 7){

  mx2[i,] <- mx[rows[i],]

}
mx2
```

Appendix 15: The 50 SWEMWBS MWB states included in the C-TTO valuation tasks

| Block | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 |
|-------|--------|--------|--------|--------|--------|--------|--------|
| 1 | 1 | 1 | 3 | 1 | 4 | 5 | 2 |
| 1 | 3 | 4 | 5 | 2 | 5 | 4 | 4 |
| 1 | 1 | 4 | 1 | 5 | 1 | 2 | 3 |
| 1 | 2 | 2 | 4 | 5 | 5 | 1 | 4 |
| 1 | 4 | 5 | 3 | 1 | 1 | 3 | 1 |
| 1 | 5 | 1 | 1 | 3 | 3 | 2 | 4 |
| 1 | 5 | 5 | 5 | 4 | 5 | 5 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 4 | 4 | 2 | 2 | 1 | 1 | 3 |
| 2 | 5 | 5 | 3 | 4 | 3 | 1 | 4 |
| 2 | 2 | 3 | 1 | 5 | 1 | 3 | 5 |
| 2 | 2 | 1 | 5 | 3 | 5 | 4 | 5 |
| 2 | 3 | 3 | 5 | 3 | 5 | 3 | 1 |
| 2 | 3 | 1 | 2 | 1 | 1 | 4 | 1 |
| 2 | 5 | 5 | 5 | 5 | 5 | 5 | 4 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 4 | 1 | 3 | 3 | 3 | 4 | 4 |
| 3 | 3 | 4 | 5 | 4 | 2 | 4 | 3 |
| 3 | 1 | 5 | 1 | 2 | 2 | 2 | 3 |
| 3 | 3 | 2 | 4 | 4 | 4 | 1 | 2 |
| 3 | 1 | 2 | 1 | 3 | 1 | 2 | 4 |
| 3 | 4 | 1 | 3 | 1 | 5 | 4 | 2 |
| 3 | 4 | 5 | 5 | 5 | 5 | 5 | 5 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 3 | 5 | 2 | 4 | 2 | 3 | 2 |
| 4 | 4 | 2 | 3 | 2 | 3 | 1 | 5 |
| 4 | 2 | 4 | 4 | 3 | 4 | 2 | 3 |
| 4 | 1 | 1 | 2 | 2 | 2 | 4 | 2 |
| 4 | 4 | 1 | 4 | 4 | 3 | 4 | 2 |
| 4 | 3 | 5 | 3 | 1 | 4 | 3 | 4 |
| 4 | 5 | 4 | 5 | 5 | 5 | 5 | 5 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5 | 1 | 4 | 3 | 5 | 4 | 5 | 5 |
| 5 | 2 | 2 | 3 | 3 | 4 | 2 | 3 |
| 5 | 5 | 2 | 5 | 4 | 1 | 1 | 3 |
| 5 | 5 | 2 | 1 | 2 | 1 | 2 | 5 |
| 5 | 3 | 4 | 1 | 1 | 4 | 4 | 1 |
| 5 | 1 | 3 | 4 | 3 | 5 | 1 | 2 |
| 5 | 5 | 5 | 5 | 5 | 5 | 4 | 5 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 6 | 4 | 1 | 4 | 2 | 2 | 3 | 1 |
| 6 | 2 | 3 | 5 | 4 | 3 | 1 | 2 |
| 6 | 3 | 3 | 2 | 1 | 3 | 3 | 4 |
| 6 | 4 | 4 | 2 | 3 | 4 | 5 | 4 |
| 6 | 1 | 1 | 2 | 5 | 2 | 2 | 1 |
| 6 | 4 | 5 | 2 | 2 | 3 | 1 | 5 |
| 6 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7 | 2 | 3 | 4 | 2 | 2 | 4 | 5 |
| 7 | 2 | 2 | 2 | 3 | 1 | 3 | 2 |
| 7 | 5 | 3 | 2 | 4 | 2 | 1 | 3 |
| 7 | 2 | 3 | 3 | 3 | 3 | 3 | 5 |
| 7 | 1 | 5 | 4 | 1 | 5 | 2 | 1 |
| 7 | 5 | 1 | 2 | 5 | 5 | 4 | 2 |
| 7 | 5 | 5 | 5 | 5 | 4 | 5 | 5 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Appendix 16: An additional version of C-TTO practice example

| | |
|---|--|
| Designated mental well-being state | You have just been for a check up with your local GP surgery who has told you that you have somewhat raised blood pressure and possible signs of diabetes. The surgery has now just telephoned to say that your blood tests have shown that a very high cholesterol and confirmed a diagnosis of diabetes. As a result, you will have to start taking pills to reduce your risk of heart disease and control your blood sugar, and also go onto a special diabetic diet. Your results mean you are at higher risk of heart disease in the future than the general population. As a result, you are very worried about your health and how you are going to manage your new diet. |
| A mental well-being state which is much higher than the designated MWB state | Now imagine that the surgery just contacted you to say that your blood tests for cholesterol and blood sugar levels are normal so you do not have to take any pills, and your risk of heart diseases is typical for a healthy person of your age. As a result, you feel greatly relieved and both happy and relaxed. |
| A mental well-being state which is much lower than the designated MWB state | Now imagine that the results from your check up showed you had diabetes and very high cholesterol and also that you have had to move house recently and have lost contact with your close friends. You feel very worried about your health, and also feel lonely and isolated because you have no-one to talk through your problems with, or to do activities you enjoy which might take your mind off your problems. |

Appendix 17: An algorithm to explore potential highly uncommon reported SWEMWBS states

It was acknowledged that there are two datasets with wide national SWEMWBS data. The first one was the Understanding Society (United Kingdom Household Longitudinal Study), which collected nine waves of data from members of the U.K. households between 2009 and 2018. SWEMWBS data was collected within waves 1, 4 and 7 of the main survey. After the elimination of missing and inapplicable data, there were 114,940 (Wave 1 = 38395; Wave 4 = 39062; Wave 7 = 37483) valid responses. The second one was the Health Survey for England, which collected annual national health data starting from 1991. The WEMWBS questionnaire was incorporated between 2010 and 2016 and there were 49081 valid data in total across all years (2010 = 7163; 2011 = 7196; 2012 = 5033; 2013 = 7777; 2014 = 7014; 2015 = 7897; 2016 = 7001). As not all items within the questionnaire were of interest, the responses of the 7 items of the SWEMWBS descriptive system were extracted from the WEMWBS descriptive system. The datasets from Understanding Society and Health Survey for England were then pooled together, resulting in a total number of 164021 SWEMWBS responses.

To investigate the response pattern of the SWEMWBS responses for identifying potential implausible states, the following step-by-step algorithm was proposed and applied to the 164,021 responses.

- 1) Identify the X most reported states.
- 2) For each of the 78,125 combinations of the SWEMWBS state,
 - a) Calculate the sum of the absolute level distances across attributes between the state and each of the X states identified in step 1. For example, the sum of the absolute level distances between the state 3333333 and state 1234543 would be $2+1+0+1+2+1+0 = 7$
 - b) Take the minimum of the sum of the absolute level distances across attributes with each of the X states. Let this value be D_i , in which i lies between 1 and 78125.
- 3) Identify and exclude the Y states with the highest values for D_i .

The frequency pattern and proportion among all of the 164,021 responses were computed. There were totally 12,801 reported states (frequency >0) and 65,324 states were

unreported (frequency =0). A selection of the top 10 responses is reported in the Appendix 17.1 below:

Appendix 17.1: The frequency and proportion of the top 10 responses

| SWEMWBS response | Frequency | Proportion (in %) |
|------------------|-----------|-------------------|
| 4444444 | 11161 | 6.80 |
| 3333333 | 6112 | 3.73 |
| 5555555 | 3738 | 2.28 |
| 3444444 | 2853 | 1.74 |
| 4434444 | 2806 | 1.71 |
| 4444445 | 1705 | 1.04 |
| 3333334 | 1504 | 0.92 |
| 3344444 | 1414 | 0.86 |
| 3434444 | 1402 | 0.85 |
| 4444434 | 1358 | 0.83 |

The table above shows that the most reported state was 4444444, which appeared 11,161 times and occupied 6.8% of the total number of responses. Since there was no official guidance regarding the threshold of values X and Y, X was defined as those states with the proportion greater than 0.05%. There were 301 states fitting this criteria (i.e. X = 301) and the lowest frequency of the state that met this criteria was 83 times. The sum of the absolute level distances across attributes between each of the 78,125 states and each of the 301 states were calculated and the corresponding D values were derived. The derived D values for each of the 78,125 states were ranged from 0 to 12. The higher the D value of a particular state, the larger would be its deviation from the 301 commonly reported states, and the higher would be the possibility that the state was uncommon or implausible for participants. It was decided that states with a D value greater than or equal to 10 should be avoided in the set of choice tasks given to the participants in the quantitative phase, as a mean to reduce the chance of encountering unearthly states during the valuation exercise.

Based on this criteria, 658 states (Y = 658) with D values between 10 and 12 have been identified. These states are listed below in the Appendix 17.2.

Appendix 17.2: States with D values between 10 and 12

| D value | State | | | | | |
|----------------|--------------|---------|---------|---------|---------|---------|
| 12 | 5515111 | 5155111 | 1555111 | 5151511 | 1551511 | 5115511 |
| | 5151151 | 1551151 | 5115151 | 1515151 | 5111551 | 5151115 |
| 11 | 5551111 | 5514111 | 5154111 | 1554111 | 5415111 | 4515111 |
| | 5525111 | 5145111 | 1545111 | 4155111 | 5255111 | 5355111 |
| | 1455111 | 2555111 | 5515211 | 5155211 | 1555211 | 5151411 |
| | 1551411 | 5115411 | 5511511 | 5141511 | 1541511 | 4151511 |
| | 5251511 | 5351511 | 1451511 | 2551511 | 5152511 | 1552511 |
| | 5114511 | 4115511 | 5215511 | 1515511 | 5125511 | 1155511 |
| | 5515121 | 5155121 | 1555121 | 5151521 | 1551521 | 5115521 |
| | 5155131 | 5151531 | 5151141 | 1551141 | 5115141 | 1515141 |
| | 5111541 | 5511151 | 5141151 | 1541151 | 4151151 | 5251151 |
| | 1451151 | 2551151 | 5152151 | 1552151 | 5114151 | 1514151 |
| | 4115151 | 5215151 | 1415151 | 2515151 | 5125151 | 1525151 |
| | 1155151 | 5155151 | 5151251 | 1551251 | 5115251 | 1515251 |
| | 5111451 | 4111551 | 5211551 | 1511551 | 5121551 | 1151551 |
| | 5151551 | 5112551 | 5515112 | 5155112 | 1555112 | 5151512 |
| | 1551512 | 5115512 | 5151152 | 1551152 | 5115152 | 1515152 |
| | 5111552 | 5151114 | 5511115 | 5141115 | 4151115 | 5251115 |
| 1551115 | 5152115 | 5115115 | 1515115 | 1155115 | 5151215 | |
| 5111515 | 5151125 | 1151155 | | | | |
| 10 | 5531111 | 5541111 | 5351111 | 5451111 | 3551111 | 4551111 |
| | 5552111 | 5513111 | 5153111 | 1553111 | 5414111 | 4514111 |
| | 5524111 | 5144111 | 1544111 | 4154111 | 5254111 | 5354111 |
| | 1454111 | 2554111 | 5315111 | 4415111 | 3515111 | 5425111 |
| | 4525111 | 5135111 | 1535111 | 5535111 | 4145111 | 5245111 |
| | 5345111 | 1445111 | 2545111 | 5545111 | 3155111 | 4255111 |
| | 1355111 | 4355111 | 2455111 | 5455111 | 3555111 | 4555111 |
| | 5555111 | 5551211 | 5514211 | 5154211 | 1554211 | 5415211 |
| | 4515211 | 5525211 | 5145211 | 1545211 | 4155211 | 5255211 |

| | | | | | |
|---------|---------|---------|---------|---------|---------|
| 5355211 | 1455211 | 2555211 | 5511311 | 5151311 | 1551311 |
| 5115311 | 5515311 | 5155311 | 1555311 | 5511411 | 5141411 |
| 1541411 | 4151411 | 5251411 | 5351411 | 1451411 | 2551411 |
| 5152411 | 1552411 | 5114411 | 4115411 | 5215411 | 1515411 |
| 5125411 | 1155411 | 5155411 | 5311511 | 5411511 | 3511511 |
| 4511511 | 5521511 | 5131511 | 1531511 | 4141511 | 5241511 |
| 5341511 | 1441511 | 2541511 | 5541511 | 3151511 | 4251511 |
| 1351511 | 4351511 | 2451511 | 5451511 | 3551511 | 4551511 |
| 5551511 | 5512511 | 5142511 | 1542511 | 4152511 | 5252511 |
| 5352511 | 1452511 | 2552511 | 5113511 | 5153511 | 1553511 |
| 4114511 | 5214511 | 1514511 | 5124511 | 1154511 | 5154511 |
| 3115511 | 4215511 | 5315511 | 1415511 | 2515511 | 5515511 |
| 4125511 | 5225511 | 1525511 | 5135511 | 1145511 | 5145511 |
| 2155511 | 4155511 | 5155511 | 1255511 | 5551121 | 5514121 |
| 5154121 | 1554121 | 5415121 | 4515121 | 5525121 | 5145121 |
| 1545121 | 4155121 | 5255121 | 5355121 | 1455121 | 2555121 |
| 5515221 | 5155221 | 1555221 | 5151421 | 1551421 | 5115421 |
| 5511521 | 5141521 | 1541521 | 4151521 | 5251521 | 5351521 |
| 1451521 | 2551521 | 5152521 | 1552521 | 5114521 | 4115521 |
| 5215521 | 1515521 | 5125521 | 1155521 | 5151131 | 1551131 |
| 5154131 | 5115131 | 1515131 | 5515131 | 5145131 | 1155131 |
| 4155131 | 5255131 | 1555131 | 5155231 | 5151431 | 5111531 |
| 5141531 | 1151531 | 4151531 | 5251531 | 1551531 | 5152531 |
| 5115531 | 5511141 | 5141141 | 1541141 | 4151141 | 5251141 |
| 1451141 | 2551141 | 5152141 | 1552141 | 5114141 | 1514141 |
| 4115141 | 5215141 | 1415141 | 2515141 | 5125141 | 1525141 |
| 1155141 | 5155141 | 5151241 | 1551241 | 5115241 | 1515241 |
| 5111441 | 4111541 | 5211541 | 1511541 | 5121541 | 1151541 |
| 5151541 | 5112541 | 1115541 | 5115541 | 5411151 | 4511151 |
| 5521151 | 5131151 | 1531151 | 4141151 | 5241151 | 1441151 |
| 2541151 | 3151151 | 4251151 | 1351151 | 5351151 | 2451151 |
| 3551151 | 5551151 | 5512151 | 5142151 | 1542151 | 4152151 |
| 5252151 | 1452151 | 2552151 | 5113151 | 1513151 | 1153151 |
| 5153151 | 1553151 | 4114151 | 5214151 | 1414151 | 2514151 |
| 5124151 | 1524151 | 1154151 | 5154151 | 3115151 | 4215151 |
| 1315151 | 5315151 | 2415151 | 5415151 | 3515151 | 4515151 |

| | | | | | |
|---------|---------|---------|---------|---------|---------|
| 5515151 | 4125151 | 5225151 | 1425151 | 2525151 | 1135151 |
| 5135151 | 1535151 | 1145151 | 5145151 | 2155151 | 3155151 |
| 4155151 | 1255151 | 5255151 | 1355151 | 5511251 | 5141251 |
| 1541251 | 4151251 | 5251251 | 1451251 | 2551251 | 5152251 |
| 1552251 | 5114251 | 1514251 | 4115251 | 5215251 | 1415251 |
| 2515251 | 5125251 | 1525251 | 1155251 | 5155251 | 5111351 |
| 1151351 | 5151351 | 1551351 | 5115351 | 1515351 | 4111451 |
| 5211451 | 1511451 | 5121451 | 1151451 | 5151451 | 5112451 |
| 1115451 | 5115451 | 3111551 | 4211551 | 5311551 | 1411551 |
| 2511551 | 5511551 | 4121551 | 5221551 | 1521551 | 1131551 |
| 5131551 | 1141551 | 5141551 | 2151551 | 3151551 | 4151551 |
| 1251551 | 5251551 | 1351551 | 4112551 | 5212551 | 1512551 |
| 5122551 | 1152551 | 5152551 | 5113551 | 1114551 | 5114551 |
| 1115551 | 5115551 | 5551112 | 5514112 | 5154112 | 1554112 |
| 5415112 | 4515112 | 5525112 | 5145112 | 1545112 | 4155112 |
| 5255112 | 5355112 | 1455112 | 2555112 | 5515212 | 5155212 |
| 1555212 | 5151412 | 1551412 | 5115412 | 5511512 | 5141512 |
| 1541512 | 4151512 | 5251512 | 5351512 | 1451512 | 2551512 |
| 5152512 | 1552512 | 5114512 | 4115512 | 5215512 | 1515512 |
| 5125512 | 1155512 | 5515122 | 5155122 | 1555122 | 5151522 |
| 1551522 | 5115522 | 5155132 | 5151532 | 5151142 | 1551142 |
| 5115142 | 1515142 | 5111542 | 5511152 | 5141152 | 1541152 |
| 4151152 | 5251152 | 1451152 | 2551152 | 5152152 | 1552152 |
| 5114152 | 1514152 | 4115152 | 5215152 | 1415152 | 2515152 |
| 5125152 | 1525152 | 1155152 | 5155152 | 5151252 | 1551252 |
| 5115252 | 1515252 | 5111452 | 4111552 | 5211552 | 1511552 |
| 5121552 | 1151552 | 5151552 | 5112552 | 5151113 | 1551113 |
| 5115113 | 5515113 | 5155113 | 1555113 | 5111513 | 5151513 |
| 1551513 | 5115513 | 5151153 | 1551153 | 5115153 | 1515153 |
| 5111553 | 5511114 | 5141114 | 4151114 | 5251114 | 1551114 |
| 5152114 | 5115114 | 1515114 | 1155114 | 5155114 | 5151214 |
| 5111514 | 1151514 | 5151514 | 5151124 | 1151154 | 1115154 |
| 5411115 | 4511115 | 5521115 | 5131115 | 4141115 | 5241115 |
| 1541115 | 3151115 | 4251115 | 5351115 | 1451115 | 2551115 |
| 5551115 | 5512115 | 5142115 | 4152115 | 5252115 | 1552115 |
| 5153115 | 5114115 | 1514115 | 1154115 | 4115115 | 5215115 |

| | | | | | | |
|--|---------|---------|---------|---------|---------|---------|
| | 1415115 | 2515115 | 5515115 | 5125115 | 1525115 | 5135115 |
| | 1145115 | 2155115 | 5155115 | 1255115 | 5511215 | 5141215 |
| | 4151215 | 5251215 | 1551215 | 5152215 | 5115215 | 1515215 |
| | 1155215 | 5151315 | 5111415 | 1151415 | 4111515 | 5211515 |
| | 1511515 | 5121515 | 1151515 | 5151515 | 5112515 | 5511125 |
| | 5141125 | 4151125 | 5251125 | 1551125 | 5152125 | 5115125 |
| | 1515125 | 1155125 | 5151225 | 5111525 | 5151135 | 5115135 |
| | 1151145 | 1115145 | 5111155 | 1511155 | 1141155 | 2151155 |
| | 5151155 | 1251155 | 1152155 | 1115155 | 5115155 | 1151255 |
| | 1111555 | | | | | |

These states were taken into account when selecting the appropriate set of choice tasks generated by the experimental designs of the C-TTO and DCE in the quantitative phase. For the C-TTO, an iteration of a set of choice tasks with a sufficiently low level-balance and without either of the Y states were chosen for the completion by participants. For the DCE, participants were allocated an iteration of a set of choice tasks with a sufficiently low D-error and without either of the Y states.

With the application of this algorithm, Appendix 17.3 shows the D values calculated from potential implausible states claimed by participants in this phase of cognitive interviews:

Appendix 17.3: The D values of implausible states claimed by participants

| Quotes | Corresponding descriptive state | Corresponding state index | D value |
|---|--|---------------------------|---------|
| <p><i>“This is an interesting health state. Coz it's quite sort of conflicted in terms of the very optimistic but then they don't feel useful at all and they're not relaxed so...”</i></p> <p><i>“but yeah again it's quite a difficult... it's quite a challenging task to do... just when you've got these really conflicting things... and then it's got lit with difficulty with it being...” [Male, 32]</i></p> <p><i>“So it's almost like you're making up your mind about things, but then... but you're also rarely dealing with problems well... it's contradictory in a way... so it's like you're able to make a decision but then your decisions are wrong, if it's associated with a problem.” [Female, 35]</i></p> | <ul style="list-style-type: none"> *often feeling optimistic about the future *rarely feeling useful *none of the time feeling relaxed *rarely dealing with problems well *some of the time thinking clearly *all of the time feeling close to other people *often able to make up my own mind about things | 4212354 | 5 |

| | | | |
|---|---|---------|---|
| <p><i>“Yup, okay so... rarely feeling optimistic is not a good thing, none of the time feeling useful is not good, all of the time feeling relaxed... so that feels counter-intuitive. [laugh]”</i></p> <p><i>“Er... all of the time feeling relaxed, so that feels strange to me that feeling relaxed but not being optimistic and not feeling useful, they seem to disagree with each other.”</i></p> <p><i>“I: Why do you think it is counter-intuitive to be none of the time feeling useful and all of the time feeling relaxed?”</i></p> <p><i>P: Yeah... if I am that relaxed all of the time, I'd probably be quite optimistic about things... me personally. And if you're feeling quite relaxed... for me again, I wouldn't be worrying about whether I felt useful or not about things, I think... that's how I would feel anyway.</i></p> <p><i>I: So is it difficult to imagine...</i></p> <p><i>P: Yes.... yeah. For me that's hard to... picture being relaxed all the time but then having these other issues going on.” [Female, 43]</i></p> | <ul style="list-style-type: none"> *rarely feeling optimistic about the future *none of the time feeling useful *all of the time feeling relaxed *some of the time dealing with problems well *all of the time thinking clearly *often feeling close to other people *all of the time able to make up my own mind about things | 2153545 | 5 |
| <p><i>“So er... this time often feeling optimistic, often feeling useful, rarely feeling relaxed, rarely dealing with problems well, none of the time</i></p> | <ul style="list-style-type: none"> *often feeling optimistic about the future *often feeling useful | 4422113 | 5 |

| | | | |
|--|--|---------|---|
| <p><i>thinking clearly, none of the time feeling close to other people, some of the time being able to make up my mind about things... so... again probably going to contradict what I've just said now, but yeah feeling... rarely feeling relaxed... yeah still feels optimistic about the future, and often feeling useful.” [Female, 43]</i></p> | <ul style="list-style-type: none"> *rarely feeling relaxed *rarely dealing with problems well *none of the time thinking clearly *none of the time feeling close to other people *some of the time able to make up my own mind about things | | |
| <p><i>“really I don't understand how you can think clearly but not deal with problems well. If you think clearly, problems should be solved.”</i></p> <p><i>“... .. You often think clearly but you... you only some of the time make up your own mind about things. So how can you be thinking clearly if you're indecisive? How can you be thinking clearly if you're unable to deal with your problems?” [Male, 32]</i></p> | <ul style="list-style-type: none"> *some of the time feeling optimistic about the future *none of the time feeling useful *often feeling relaxed *some of the time dealing with problems well *often thinking clearly *all of the time feeling close to other people *some of the time able to make up my own mind about things | 3143453 | 4 |
| <p><i>“You know you should not feel optimistic if you never feel useful.”</i></p> <p><i>“Often feel relaxed, often deal with problems... hmm... often deal with problems well despite the fact that you can't think clearly now, that is strange. And you can rarely make up your mind, now this does not</i></p> | <ul style="list-style-type: none"> *often feeling optimistic about the future *none of the time feeling useful *often feeling relaxed *often dealing with problems well *some of the time thinking clearly | 4144342 | 4 |

| | | | |
|--|---|---------|---|
| <p><i>make sense. I mean how can I only think clearly some of the time and I can't make my mind up about anything, but I can deal with problems well often! This does not make sense.” [Female, 67]</i></p> | <ul style="list-style-type: none"> *often feeling close to other people *rarely able to make up my own mind about things | | |
| <p><i>“Now the other way around and I'll start with B... so I'm optimistic... I think clearly... but I don't feel useful, I don't feel relaxed, I don't think, I don't feel close to people, and I can't make my mind up, but I feel optimistic, and I think clearly, none of this makes sense. [laugh] None of this makes sense. I can think clearly and I can feel optimistic despite the fact that I don't feel useful, I don't deal with problems, I can't make my mind up and I don't feel close to people... oh this can't work.” [Female, 67]</i></p> | <ul style="list-style-type: none"> *all of the time feeling optimistic about the future *none of the time feeling useful *rarely feeling relaxed *none of the time dealing with problems well *all of the time thinking clearly *none of the time feeling close to other people *none of the time able to make up my own mind about things | 5121511 | 9 |
| <p><i>“But you might say something like I've been dealing with problems well but I can't make up my mind. And you think... well if I can't make up my mind, how can you even start to deal with problems... it just didn't make sense.” [Female, 67]</i></p> | <ul style="list-style-type: none"> *rarely feeling optimistic about the future *often feeling useful *none of the time feeling relaxed *often dealing with problems well *rarely thinking clearly *all of the time feeling close to other people *rarely able to make up my own mind about things | 2414252 | 7 |

| | | | |
|---|--|---------|---|
| <p><i>"... again it sounds counter-intuitive to me em... because if I'm not thinking clearly, it's... difficult to see how I'm dealing with problems well." [Male, 67]</i></p> | <ul style="list-style-type: none"> *all of the time feeling optimistic about the future *rarely feeling useful *all of the time feeling relaxed *often dealing with problems well *none of the time thinking clearly *none of the time feeling close to other people *some of the time able to make up my own mind about things | 5254113 | 8 |
| <p><i>"See again... sometimes it's rarely dealing with problems well, often thinking clearly, er... seem counter-intuitive to me. They don't... really go together." [Male, 67]</i></p> | <ul style="list-style-type: none"> *some of the time feeling optimistic about the future *often feeling useful *rarely feeling relaxed *rarely dealing with problems well *often thinking clearly *rarely feeling close to other people *rarely able to make up my own mind about things | 3422422 | 4 |

I indicates interviewer; P, participant.

Interestingly, the D statistics for the states which caused concerns to participants were actually not that implausible since they were not that dissimilar from states actually reported in the national survey data. Most of the D statistics were around 5, with only one state with D-statistics of 9. In this context, there was insufficient evidence to rule out any states when running the experimental designs of C-TTO and DCE. However, it was worth noting that the selection of those choice sets which include the states with D-values of 10 or above could be avoided, with a view to minimise imagination burden. It should be stressed that I was not suggesting the deletion of those highly uncommon states as they were still extrapolated during the modelling of utility values in Chapter 6 when conducting a larger SWEMWBS valuation study.

Appendix 18: An example of the advertisement for interview recruitment


Study title: Valuation of mental well-being as measured by the Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS): A quantitative study

Are you interested in a mental well-being study? Do you want to contribute towards the development of valuation methods for mental well-being? Here is the chance!

- **Study aim:** To test approaches developed by economists for valuing health states and see how well they apply to valuing mental well-being states. Your participation will help us understand preferences for different approaches.

What will be done during the interview?

- To make trade-off between choices of imaginable life with different durations of mental well-being status.
- To look at pairs of mental well-being profiles and choose the one you prefer.
- To answer several debriefing questions regarding the difficulties of understanding the instructions and tasks.



Eligibility:

- ✓ Aged 18 or above
- ✓ Possession of the right to vote in the U.K.

Name of interviewer: Hei Hang Edmund Yiu, Ph.D. student at Warwick Medical School, University of Warwick

Interview duration: ~60 minutes

Venue: A virtual face-to-face meeting using Microsoft Teams for which a stable internet connection is required. Instruction regarding the installation of the Teams desktop app will be provided

Contact email: H.Yiu@warwick.ac.uk

Please contact me by email if you are happy to participate. You will have a chance to win the lottery of a £25 gift voucher upon completion of the interview!
I look forward to hearing from you! 😊

Appendix 19: Syntax for the DCE experimental design in Ngene

```
Design
;alts = alt1, alt2
;rows = 50
;eff = (mnl,d)
;block = 5
;model:

U(alt1) = b0[0] + b1.dummy[(u,-4,0)|(u,-3,0)|(u,-2,0)|(u,-
1,1)]*item1[1,2,3,4,5] + b2.dummy[(u,-4,0)|(u,-3,0)|(u,-2,0)|(u,-
1,1)]*item2[1,2,3,4,5] + b3.dummy[(u,-4,0)|(u,-3,0)|(u,-2,0)|(u,-
1,1)]*item3[1,2,3,4,5] + b4.dummy[(u,-4,0)|(u,-3,0)|(u,-2,0)|(u,-
1,1)]*item4[1,2,3,4,5] + b5.dummy[(u,-4,0)|(u,-3,0)|(u,-2,0)|(u,-
1,1)]*item5[1,2,3,4,5] + b6.dummy[(u,-4,0)|(u,-3,0)|(u,-2,0)|(u,-
1,1)]*item6[1,2,3,4,5] + b7.dummy[(u,-4,0)|(u,-3,0)|(u,-2,0)|(u,-
1,1)]*item7[1,2,3,4,5] /

U(alt2) = b1.dummy*item1 + b2.dummy*item2 + b3.dummy*item3 +
b4.dummy*item4 + b5.dummy*item5 + b6.dummy*item6 + b7.dummy*item7

$
```

Note: The prior uncertainty was constructed by specifying uniformly distributed Bayesian priors for all dummy variables. The negative lower-end distribution value for each dummy variable indicated the disutility from the best level (i.e. level 5). The lower (higher) the mental well-being level, the higher (lower) would be the expected disutility. It was expected that the dummy variable for level 1 (none of the time) exhibited the largest disutility. The higher-end distribution value for the dummy variable of level 4 was set as positive, to account for the fact that the sign of non-monotonic valuation (i.e. not preferring full mental well-being) was discovered in the qualitative phase.

Appendix 20: The 50 pairs of mental well-being states included in the DCE valuation tasks

| Block | Alternative 1 | | | | | | | Alternative 2 | | | | | | |
|-------|---------------|---|---|---|---|---|---|---------------|---|---|---|---|---|---|
| 1 | 4 | 4 | 3 | 5 | 1 | 1 | 3 | 2 | 3 | 1 | 4 | 2 | 5 | 5 |
| 1 | 3 | 5 | 5 | 3 | 2 | 2 | 3 | 1 | 1 | 4 | 4 | 1 | 5 | 4 |
| 1 | 5 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 3 | 3 | 1 | 4 | 3 | 3 |
| 1 | 1 | 3 | 1 | 3 | 5 | 5 | 1 | 3 | 5 | 2 | 1 | 4 | 1 | 2 |
| 1 | 3 | 1 | 4 | 4 | 5 | 5 | 1 | 4 | 2 | 1 | 3 | 3 | 4 | 4 |
| 1 | 2 | 5 | 5 | 4 | 1 | 1 | 4 | 3 | 4 | 4 | 3 | 4 | 2 | 1 |
| 1 | 4 | 1 | 4 | 2 | 1 | 4 | 1 | 1 | 2 | 2 | 3 | 5 | 3 | 3 |
| 1 | 3 | 2 | 1 | 5 | 1 | 5 | 2 | 5 | 4 | 3 | 1 | 3 | 1 | 1 |
| 1 | 1 | 1 | 5 | 4 | 3 | 1 | 3 | 2 | 2 | 2 | 1 | 4 | 2 | 4 |
| 1 | 5 | 4 | 5 | 2 | 1 | 5 | 4 | 4 | 5 | 4 | 3 | 4 | 1 | 5 |
| 2 | 5 | 5 | 1 | 2 | 4 | 3 | 1 | 3 | 3 | 3 | 3 | 3 | 4 | 2 |
| 2 | 1 | 2 | 4 | 5 | 5 | 4 | 5 | 5 | 3 | 5 | 4 | 3 | 3 | 2 |
| 2 | 2 | 4 | 5 | 1 | 5 | 1 | 4 | 5 | 5 | 3 | 2 | 4 | 3 | 2 |
| 2 | 1 | 4 | 3 | 2 | 5 | 4 | 2 | 3 | 2 | 5 | 4 | 3 | 1 | 4 |
| 2 | 1 | 3 | 2 | 2 | 5 | 2 | 5 | 2 | 1 | 3 | 5 | 2 | 5 | 1 |
| 2 | 4 | 3 | 1 | 1 | 4 | 2 | 2 | 3 | 2 | 2 | 3 | 2 | 5 | 1 |
| 2 | 3 | 1 | 5 | 3 | 4 | 5 | 5 | 4 | 3 | 2 | 4 | 3 | 3 | 4 |
| 2 | 3 | 2 | 1 | 5 | 3 | 3 | 3 | 2 | 5 | 4 | 3 | 1 | 2 | 2 |
| 2 | 4 | 2 | 4 | 2 | 4 | 1 | 2 | 5 | 3 | 3 | 5 | 1 | 4 | 1 |
| 2 | 4 | 2 | 3 | 1 | 2 | 2 | 4 | 2 | 3 | 1 | 3 | 1 | 4 | 5 |
| 3 | 1 | 5 | 2 | 2 | 3 | 4 | 4 | 5 | 1 | 5 | 1 | 5 | 3 | 2 |
| 3 | 5 | 2 | 1 | 4 | 4 | 4 | 5 | 2 | 5 | 2 | 5 | 5 | 5 | 4 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 3 | 5 | 1 | 5 | 3 |
| 3 | 4 | 3 | 3 | 4 | 5 | 3 | 1 | 5 | 4 | 1 | 3 | 3 | 5 | 3 |
| 3 | 5 | 4 | 2 | 4 | 1 | 2 | 5 | 1 | 5 | 5 | 2 | 5 | 3 | 3 |
| 3 | 3 | 5 | 3 | 3 | 1 | 3 | 5 | 2 | 4 | 2 | 2 | 5 | 4 | 3 |
| 3 | 2 | 5 | 3 | 4 | 2 | 5 | 3 | 5 | 2 | 5 | 5 | 1 | 4 | 1 |
| 3 | 4 | 3 | 4 | 1 | 2 | 5 | 3 | 3 | 5 | 3 | 4 | 5 | 1 | 2 |
| 3 | 5 | 2 | 4 | 3 | 1 | 4 | 2 | 4 | 1 | 2 | 5 | 3 | 1 | 5 |
| 3 | 2 | 1 | 5 | 5 | 4 | 2 | 4 | 4 | 5 | 2 | 1 | 2 | 5 | 5 |
| 4 | 4 | 3 | 1 | 3 | 5 | 1 | 3 | 2 | 1 | 5 | 1 | 2 | 4 | 5 |
| 4 | 5 | 1 | 3 | 3 | 4 | 1 | 4 | 2 | 2 | 4 | 5 | 1 | 2 | 5 |
| 4 | 2 | 2 | 3 | 4 | 2 | 2 | 3 | 1 | 4 | 4 | 1 | 3 | 3 | 2 |
| 4 | 2 | 2 | 3 | 2 | 3 | 3 | 5 | 1 | 1 | 2 | 5 | 4 | 4 | 1 |
| 4 | 2 | 4 | 2 | 1 | 4 | 5 | 5 | 4 | 1 | 5 | 3 | 2 | 4 | 4 |
| 4 | 1 | 5 | 1 | 4 | 2 | 4 | 2 | 4 | 4 | 3 | 2 | 1 | 1 | 5 |
| 4 | 1 | 4 | 3 | 4 | 3 | 4 | 1 | 5 | 3 | 4 | 2 | 2 | 1 | 4 |
| 4 | 3 | 1 | 2 | 1 | 4 | 4 | 4 | 1 | 2 | 4 | 4 | 3 | 2 | 5 |
| 4 | 4 | 1 | 2 | 3 | 5 | 3 | 2 | 1 | 3 | 3 | 2 | 4 | 1 | 4 |
| 4 | 2 | 3 | 5 | 5 | 4 | 3 | 1 | 3 | 1 | 4 | 2 | 5 | 2 | 4 |
| 5 | 2 | 4 | 2 | 3 | 1 | 3 | 3 | 3 | 3 | 1 | 5 | 5 | 2 | 1 |
| 5 | 3 | 4 | 4 | 5 | 2 | 1 | 5 | 2 | 5 | 5 | 4 | 4 | 2 | 3 |
| 5 | 1 | 1 | 2 | 2 | 3 | 2 | 5 | 5 | 2 | 1 | 1 | 5 | 1 | 2 |
| 5 | 2 | 1 | 1 | 5 | 2 | 3 | 4 | 1 | 2 | 5 | 2 | 1 | 5 | 3 |
| 5 | 3 | 3 | 4 | 1 | 3 | 5 | 1 | 1 | 4 | 1 | 5 | 2 | 2 | 5 |
| 5 | 5 | 5 | 4 | 1 | 5 | 2 | 3 | 4 | 4 | 5 | 2 | 2 | 5 | 2 |
| 5 | 5 | 2 | 2 | 5 | 3 | 4 | 2 | 4 | 4 | 1 | 4 | 4 | 2 | 1 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1 | 3 | 4 | 5 | 1 | 2 | 2 | 4 | 1 | 1 | 2 | 5 | 4 | 1 |
| 5 | 5 | 3 | 2 | 3 | 2 | 1 | 1 | 3 | 5 | 1 | 1 | 1 | 3 | 3 |
| 5 | 4 | 5 | 5 | 1 | 3 | 5 | 2 | 3 | 4 | 4 | 4 | 2 | 3 | 3 |

Appendix 21: Code for the C-TTO experimental design in R

```
rm(list=ls())

# load files with states format should be: one column per dimension;
# header row with dimension name; each row includes one state

# TTO_fixed contains the 7 mildest states and the worst state

# TTO_all contains all states except the worst and the 7 mildest

TTO_fixed <- read.csv("D:/Warwick/PhD/Preference
tariff/Thesis/Experimental design/Composite TTO/TTO_fixed_states.csv")

TTO_all <- read.csv("D:/Warwick/PhD/Preference
tariff/Thesis/Experimental design/Composite TTO/TTO_all_states.csv")

num_fix <- 8
num_var <- 56
num_all <- 78117

lvldistmat <- array(0,dim=c(10,7))

# loop
lvlbalancecheck <- 10000000000
lvldistmatbest <- array(0,dim=c(10,7))
TTOrandbest <- rep(NA,56)

# number of iterations is still 10,000.
for (m in 1:10000){
  EQ <- array(0,dim=c(num_fix+num_var,7))
  EQlvmat <- array(0,dim=c(5,7))
  # select random subset of "num_var" states from TTO_all which
  # includes "num_all" states in total
  TTO_rand <- TTO_all[sample(num_all,num_var,F),]
  dim(TTO_rand)
  # combine the sampled and fixed states into a single design
  TTO <- rbind(TTO_fixed,TTO_rand)
  # check design for level balance
  for(i in 1:7){
    EQ[,i] <- TTO[,i]
  }
  for (j in 1:7) {
    for (k in 1:5) {
      EQlvmat[k,j]<-sum(EQ[,j]==k)
```

```

    }
  }

  for (j in 1:7) {
    lvldistmat[1,j]<-(EQlvlmat[1,j]-EQlvlmat[2,j])^2
    lvldistmat[2,j]<-(EQlvlmat[1,j]-EQlvlmat[3,j])^2
    lvldistmat[3,j]<-(EQlvlmat[1,j]-EQlvlmat[4,j])^2
    lvldistmat[4,j]<-(EQlvlmat[1,j]-EQlvlmat[5,j])^2

    lvldistmat[5,j]<-(EQlvlmat[2,j]-EQlvlmat[3,j])^2
    lvldistmat[6,j]<-(EQlvlmat[2,j]-EQlvlmat[4,j])^2
    lvldistmat[7,j]<-(EQlvlmat[2,j]-EQlvlmat[5,j])^2

    lvldistmat[8,j]<-(EQlvlmat[3,j]-EQlvlmat[4,j])^2
    lvldistmat[9,j]<-(EQlvlmat[3,j]-EQlvlmat[5,j])^2

    lvldistmat[10,j]<-(EQlvlmat[4,j]-EQlvlmat[5,j])^2
  }
  lvlbalcheck2 <- sqrt(sum(lvldistmat[,j]))
  if (lvlbalcheck2<lvlbalcheck){
    lvlbalcheck <- lvlbalcheck2
    lvldistmatbest <- lvldistmat
    TTo randbest <- TTo_rand
  }
  ###lvlbalcheck
}
head(lvldistmatbest)
lvlbalcheck
##### Blocking
library("AlgDesign")
des <- TTo randbest
## number of pairs per respondent
npairs <- 8
## number of blocks

```

```
nblocks <- 7
desBlock<-optBlock(~.,des,c(rep(npairs,nblocks)))
print(desBlock$Blocks)
TTO_blocked<-desBlock$Blocks
TTO_blocked
rows<-sample.int(7,7,replace=FALSE)
for ( i in 1: 7){
mx <- rbind(mx,TTO_fixed[rows[i],])
}
mx
mx <- TTO_fixed[2:8,]
mx2 <- mx
for ( i in 1: 7){
mx2[i,] <- mx[rows[i],]
}
mx2
```

Appendix 22: The 64 SWEMWBS mental well-being states included in the C-TTO valuation tasks

| Block | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 |
|-------|--------|--------|--------|--------|--------|--------|--------|
| 1 | 3 | 2 | 1 | 5 | 2 | 2 | 1 |
| 1 | 2 | 4 | 3 | 2 | 3 | 3 | 5 |
| 1 | 4 | 3 | 5 | 2 | 2 | 3 | 2 |
| 1 | 2 | 1 | 4 | 5 | 5 | 4 | 2 |
| 1 | 1 | 4 | 3 | 3 | 4 | 5 | 3 |
| 1 | 3 | 4 | 4 | 2 | 2 | 2 | 2 |
| 1 | 4 | 1 | 2 | 1 | 5 | 5 | 4 |
| 1 | 5 | 4 | 1 | 4 | 2 | 1 | 3 |
| 1 | 5 | 5 | 5 | 5 | 4 | 5 | 5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 3 | 5 | 3 | 1 | 1 | 4 | 2 |
| 2 | 1 | 4 | 1 | 5 | 1 | 4 | 4 |
| 2 | 4 | 3 | 3 | 3 | 5 | 2 | 5 |
| 2 | 3 | 2 | 2 | 1 | 2 | 1 | 1 |
| 2 | 4 | 2 | 5 | 2 | 4 | 5 | 2 |
| 2 | 1 | 3 | 2 | 2 | 1 | 1 | 3 |
| 2 | 2 | 4 | 4 | 4 | 4 | 3 | 2 |
| 2 | 5 | 1 | 2 | 3 | 5 | 5 | 4 |
| 2 | 5 | 5 | 5 | 5 | 5 | 5 | 4 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 4 | 1 | 1 | 2 | 2 | 5 | 5 |
| 3 | 5 | 1 | 4 | 2 | 2 | 5 | 1 |
| 3 | 3 | 2 | 4 | 3 | 4 | 1 | 3 |
| 3 | 5 | 4 | 4 | 5 | 3 | 3 | 5 |
| 3 | 3 | 5 | 2 | 2 | 1 | 3 | 4 |
| 3 | 2 | 5 | 5 | 4 | 5 | 2 | 1 |
| 3 | 1 | 4 | 1 | 2 | 4 | 5 | 2 |
| 3 | 1 | 1 | 1 | 2 | 2 | 1 | 1 |
| 3 | 5 | 4 | 5 | 5 | 5 | 5 | 5 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 4 | 5 | 1 | 3 | 2 | 4 | 5 |
| 4 | 5 | 4 | 3 | 4 | 3 | 1 | 2 |
| 4 | 3 | 3 | 3 | 4 | 3 | 3 | 2 |
| 4 | 4 | 1 | 5 | 4 | 1 | 4 | 4 |
| 4 | 1 | 3 | 4 | 2 | 3 | 1 | 2 |
| 4 | 3 | 1 | 4 | 4 | 4 | 4 | 2 |
| 4 | 1 | 4 | 1 | 1 | 4 | 1 | 1 |
| 4 | 1 | 3 | 2 | 1 | 3 | 5 | 4 |
| 4 | 5 | 5 | 4 | 5 | 5 | 5 | 5 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | 3 | 1 | 3 | 3 | 4 | 5 | 5 |
| 5 | 2 | 1 | 2 | 2 | 1 | 3 | 3 |
| 5 | 3 | 5 | 3 | 1 | 2 | 2 | 3 |
| 5 | 2 | 2 | 5 | 1 | 1 | 2 | 2 |
| 5 | 2 | 5 | 2 | 1 | 1 | 1 | 2 |
| 5 | 3 | 5 | 2 | 5 | 5 | 3 | 3 |
| 5 | 4 | 2 | 2 | 4 | 5 | 5 | 1 |
| 5 | 3 | 2 | 3 | 5 | 4 | 4 | 3 |
| 5 | 5 | 5 | 5 | 4 | 5 | 5 | 5 |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | 4 | 4 | 5 | 4 | 1 | 4 | 4 |
| 6 | 1 | 5 | 2 | 3 | 4 | 4 | 2 |
| 6 | 2 | 3 | 4 | 3 | 5 | 5 | 1 |
| 6 | 1 | 2 | 4 | 4 | 4 | 1 | 4 |
| 6 | 1 | 2 | 3 | 3 | 3 | 2 | 3 |
| 6 | 5 | 2 | 1 | 1 | 1 | 4 | 1 |
| 6 | 3 | 3 | 5 | 4 | 2 | 3 | 2 |
| 6 | 5 | 2 | 1 | 1 | 4 | 2 | 4 |
| 6 | 5 | 5 | 5 | 5 | 5 | 4 | 5 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 5 | 1 | 1 | 4 | 3 |
| 7 | 1 | 4 | 4 | 1 | 3 | 2 | 5 |
| 7 | 2 | 5 | 2 | 1 | 2 | 1 | 4 |
| 7 | 3 | 1 | 1 | 5 | 4 | 1 | 3 |
| 7 | 5 | 1 | 2 | 4 | 3 | 5 | 4 |
| 7 | 2 | 3 | 2 | 3 | 2 | 4 | 1 |
| 7 | 4 | 2 | 5 | 5 | 4 | 3 | 1 |
| 7 | 4 | 5 | 1 | 2 | 5 | 4 | 1 |
| 7 | 4 | 5 | 5 | 5 | 5 | 5 | 5 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Appendix 23: Facebook advertising statistics

| Number | Date | Target | Women | Men | Difference between women and men (%) | Duration (day) | Reach | Link clicks | Cost per link click |
|---------------|-------------|---------------|--------------|------------|---|-----------------------|--------------|--------------------|----------------------------|
| 1 | 21 Jan 2021 | Women and Men | 79.50% | 20.50% | 59.00% | 2 | 2,510 | 59 | £0.17 |
| 2 | 26 Jan 2021 | Women and Men | 85.70% | 14.30% | 71.40% | 7 | 25,559 | 835 | £0.09 |
| 3 | 18 Feb 2021 | Women and Men | 81.80% | 18.20% | 63.60% | 2 | 3,429 | 102 | £0.13 |
| 4 | 23 Feb 2021 | Women and Men | 72.60% | 27.40% | 45.20% | 2 | 14,197 | 217 | £0.14 |
| 5 | 03 Mar 2021 | Women and Men | 86.70% | 13.30% | 73.40% | 2 | 9,519 | 171 | £0.14 |
| 6 | 09 Mar 2021 | Women and Men | 85.10% | 14.90% | 70.20% | 2 | 8,544 | 203 | £0.14 |
| 7 | 15 Mar 2021 | Women and Men | 61.20% | 38.80% | 22.40% | 3 | 25,680 | 264 | £0.17 |

| | | | | | | | | | |
|----|-------------|---------------|--------|---------|--------|---|--------|-----|-------|
| 8 | 20 Mar 2021 | Women and Men | 79.10% | 20.90% | 58.20% | 2 | 15,132 | 115 | £0.26 |
| 9 | 22 Mar 2021 | Women and Men | 73.50% | 26.50% | 47.00% | 3 | 21,199 | 191 | £0.24 |
| 10 | 19 Apr 2021 | Women and Men | 84.50% | 15.50% | 69.00% | 4 | 21,276 | 463 | £0.13 |
| 11 | 25 Apr 2021 | Women and Men | 90.40% | 9.60% | 80.80% | 3 | 19,758 | 342 | £0.13 |
| 12 | 02 May 2021 | Women and Men | 90.10% | 9.90% | 80.20% | 3 | 22,190 | 305 | £0.15 |
| 13 | 09 May 2021 | Women and Men | 90.50% | 9.50% | 81.00% | 4 | 23,606 | 344 | £0.17 |
| 14 | 15 May 2021 | Women and Men | 91.70% | 8.30% | 83.40% | 4 | 28,200 | 400 | £0.15 |
| 15 | 21 May 2021 | Women and Men | 91.60% | 8.40% | 83.20% | 3 | 20,128 | 289 | £0.16 |
| 16 | 02 Jun 2021 | Men | 0.00% | 100.00% | NA | 4 | 17,106 | 186 | £0.32 |
| 17 | 07 Jun 2021 | Men | 0.00% | 100.00% | NA | 4 | 15,420 | 141 | £0.41 |
| 18 | 10 Jun 2021 | Women and Men | 94.30% | 5.70% | 88.60% | 4 | 20,100 | 426 | £0.14 |

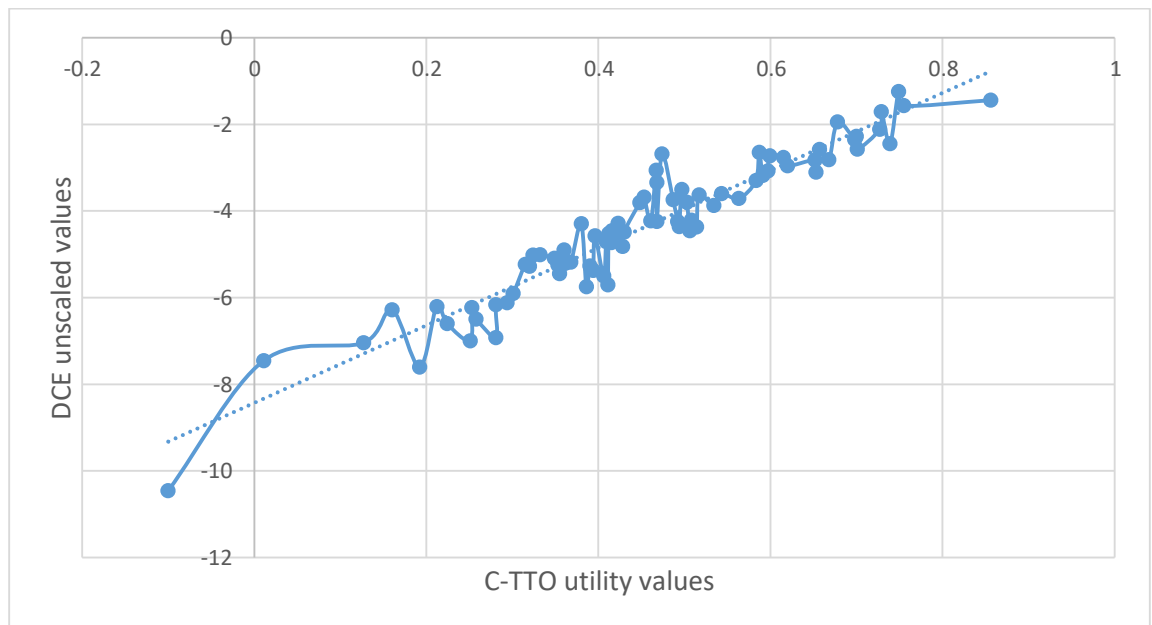
| | | | | | | | | | |
|----|-------------|---------------|--------|--------|--------|---|--------|-----|-------|
| 19 | 14 Jun 2021 | Women and Men | 94.00% | 6.00% | 88.00% | 4 | 24,080 | 376 | £0.16 |
| 20 | 21 Jun 2021 | Women and Men | 91.70% | 8.30% | 83.40% | 2 | 14,192 | 185 | £0.16 |
| 21 | 04 Jul 2021 | Women and Men | 93.70% | 6.30% | 87.40% | 3 | 28,368 | 444 | £0.10 |
| 22 | 11 Jul 2021 | Women and Men | 95.10% | 4.90% | 90.20% | 2 | 14,636 | 213 | £0.09 |
| 23 | 17 Jul 2021 | Women and Men | 89.10% | 10.90% | 78.20% | 4 | 24,905 | 404 | £0.14 |

NA indicates not applicable.

Appendix 24: The EuroQol hybrid model

The assumption behind this hybrid model required the proportionality between DCE unscaled values and C-TTO utility values. A graphical plot below shows this relationship:

Appendix 24.1: A graphical relationship between DCE unscaled values and C-TTO utility values



The selection of the C-TTO utility values derived from Model 1A and the DCE unscaled values produced by Model 2A are presented in this figure, ordered by the C-TTO utility values for the horizontal axis. As a general trend line could be fit into the C-TTO and DCE data, the proportional relationship was confirmed and the hybrid model could be safely estimated.

The hybrid coefficients generated using the “hyreg” command in Stata are presented in Appendix 24.2 below.

Appendix 24.2: Modelling result of the EuroQol hybrid model

| | b (R.SE) | p |
|-------------|----------------------|-------|
| optimistic1 | -0.154*** (0.036) | 0.000 |
| optimistic2 | -0.136*** (0.034) | 0.000 |
| optimistic3 | -0.060* | 0.092 |

| | | |
|------------------|------------------|-------|
| | (0.036) | |
| optimistic4 | -0.072** | 0.019 |
| | (0.031) | |
| useful1 | -0.143*** | 0.000 |
| | (0.026) | |
| useful2 | -0.118*** | 0.001 |
| | (0.034) | |
| useful3 | -0.029 | 0.420 |
| | (0.035) | |
| useful4 | -0.035 | 0.245 |
| | (0.030) | |
| relaxed1 | -0.169*** | 0.000 |
| | (0.035) | |
| relaxed2 | -0.186*** | 0.000 |
| | (0.036) | |
| relaxed3 | -0.086** | 0.023 |
| | (0.038) | |
| relaxed4 | -0.066* | 0.055 |
| | (0.034) | |
| dealingproblems1 | -0.100*** | 0.006 |
| | (0.037) | |
| dealingproblems2 | -0.105*** | 0.001 |
| | (0.033) | |
| dealingproblems3 | -0.073** | 0.040 |
| | (0.036) | |
| dealingproblems4 | -0.018 | 0.497 |
| | (0.027) | |
| thinkingclearly1 | -0.205*** | 0.000 |
| | (0.038) | |
| thinkingclearly2 | -0.108*** | 0.001 |
| | (0.032) | |
| thinkingclearly3 | -0.051 | 0.193 |
| | (0.039) | |
| thinkingclearly4 | -0.015 | 0.623 |
| | (0.030) | |
| closetopeople1 | -0.164*** | 0.000 |

| | | |
|----------------|--------------|-------|
| | (0.033) | |
| closetopeople2 | -0.163*** | 0.000 |
| | (0.038) | |
| closetopeople3 | -0.072** | 0.028 |
| | (0.033) | |
| closetopeople4 | 0.002 | 0.931 |
| | (0.029) | |
| makeupownmind1 | -0.168*** | 0.000 |
| | (0.034) | |
| makeupownmind2 | -0.125*** | 0.000 |
| | (0.032) | |
| makeupownmind3 | -0.086** | 0.020 |
| | (0.037) | |
| makeupownmind4 | -0.022 | 0.473 |
| | (0.030) | |
| constant | 1.175*** | 0.000 |
| | (0.033) | |

Intheta

| | | |
|----------|-----------|-------|
| constant | -3.290*** | 0.002 |
| | (1.066) | |

AIC 5441.010

BIC 5817.295

N 4349.000

Number of statistically significant main effects parameters at 5% level 19

Number of statistically significant main effects parameters at 10% level 21

Notes: *** significant at 1%, **significant at 5%, * significant at 10%

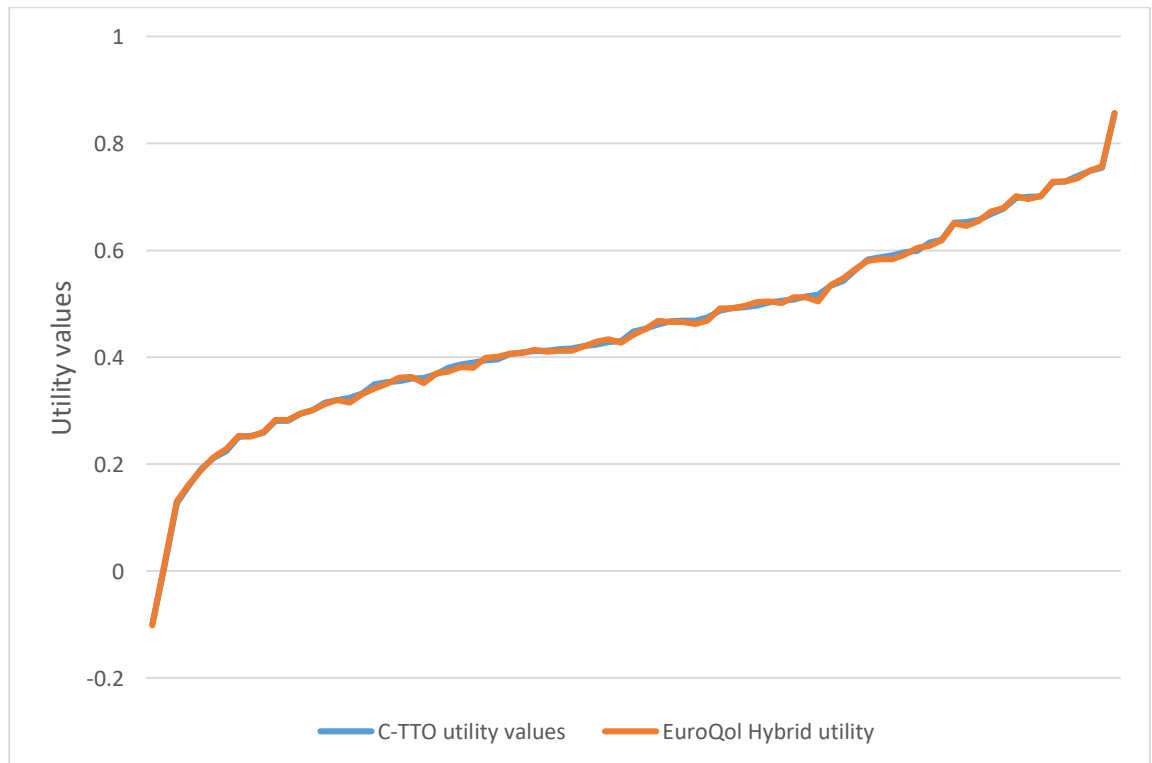
Robust standard errors are presented in parentheses.

Potentially logical inconsistent coefficients are highlighted in bold.

AIC indicates Akaike Information Criterion; BIC, Bayesian Information Criterion; N, number of observations

However, it was interesting to realise that the valuation set generated by these coefficients was roughly the same as the valuation set generated by the Model 1A. In other words, contrary to my expectation, the result of the hybrid model was nearly perfectly biased towards the C-TTO models. This finding is illustrated by a graphical plot between the C-TTO utility values and hybrid utility values in Appendix 24.3 below.

Appendix 24.3: A graphical relationship between selected C-TTO utility values and the EuroQol hybrid utility values, ordered by the C-TTO utility values



This graph obviously suggests that the hybrid result did not really take the DCE modelling result into consideration.

Furthermore, the rescaling θ (0.0372) produced by the hybrid model was not informative. For example, applying the rescaling formula (TTO coefficient = θ *rescaled DCE coefficient), the rescaled DCE utility value for the state 1111111 is as follows:

$$1 - \left(-\frac{0.153}{0.03724}\right) - \left(-\frac{0.144}{0.03724}\right) - \left(-\frac{0.168}{0.03724}\right) - \left(-\frac{0.097}{0.03724}\right) - \left(-\frac{0.204}{0.03724}\right) - \left(-\frac{0.169}{0.03724}\right) - \left(-\frac{0.165}{0.03724}\right) = -28.54$$

The utility value was unexplainedly large and no conclusive evidence could be made to justify this.

I approached a EuroQol member for consulting this issue. He was one of the developers of this “hyreg” command and he was experienced in applying this hybrid method to model the C-TTO and DCE data, as indicated by the possession of relevant publications (e.g. Ramos-Goñi *et al.* (2016); Ramos-Goni *et al.* (2017b).) He suggested different debugging methods in Stata and diagnostic checks to test the validity of my C-TTO and DCE models. He also provided guidance on formatting the data sheet and constructing the relevant command for running the hybrid model in Stata. However, even though he confirmed the validity of my Stata commands and robustness of my C-TTO and DCE models, we could not figure out the reason of deriving this unfavourable hybrid result in my case. Unfortunately, taking into account the expert’s opinion, the result of this hybrid model was not adopted in this thesis.

Appendix 25: Sensitivity analysis of the C-TTO data

Model 4: Main effects

| | b (R.SE) | p |
|------------------|-----------------------------|-------|
| optimistic1 | -0.159*** (0.034) | 0.000 |
| optimistic2 | -0.146*** (0.032) | 0.000 |
| optimistic3 | -0.076** (0.034) | 0.023 |
| optimistic4 | -0.088*** (0.029) | 0.003 |
| useful1 | -0.146*** (0.025) | 0.000 |
| useful2 | -0.112*** (0.032) | 0.001 |
| useful3 | -0.027 (0.034) | 0.430 |
| useful4 | -0.032 (0.029) | 0.272 |
| relaxed1 | -0.159*** (0.033) | 0.000 |
| relaxed2 | -0.172*** (0.034) | 0.000 |
| relaxed3 | -0.080** (0.036) | 0.028 |
| relaxed4 | -0.061* (0.033) | 0.067 |
| dealingproblems1 | -0.092*** (0.035) | 0.009 |
| dealingproblems2 | -0.105*** (0.032) | 0.001 |
| dealingproblems3 | -0.068** | 0.040 |

| | | |
|------------------|--------------|-------|
| | (0.033) | |
| dealingproblems4 | -0.014 | 0.582 |
| | (0.026) | |
| thinkingclearly1 | -0.197*** | 0.000 |
| | (0.037) | |
| thinkingclearly2 | -0.105*** | 0.001 |
| | (0.030) | |
| thinkingclearly3 | -0.058 | 0.118 |
| | (0.037) | |
| thinkingclearly4 | -0.011 | 0.716 |
| | (0.030) | |
| closetopeople1 | -0.173*** | 0.000 |
| | (0.030) | |
| closetopeople2 | -0.169*** | 0.000 |
| | (0.034) | |
| closetopeople3 | -0.062** | 0.049 |
| | (0.031) | |
| closetopeople4 | 0.007 | 0.814 |
| | (0.028) | |
| makeupownmind1 | -0.167*** | 0.000 |
| | (0.033) | |
| makeupownmind2 | -0.124*** | 0.000 |
| | (0.030) | |
| makeupownmind3 | -0.085** | 0.014 |
| | (0.035) | |
| makeupownmind4 | -0.027 | 0.346 |
| | (0.029) | |
| _cons | 1.176*** | 0.000 |
| | (0.032) | |
| <hr/> | | |
| AIC | 2450.295 | |
| BIC | 2781.979 | |
| N | 2250 | |

Number of statistically significant main effects parameters at 5% level 20

Number of statistically 21
significant main effects
parameters at 10% level

Notes: *** significant at 1%, **significant at 5%, * significant at 10%

Robust standard errors are presented in parentheses.

Potentially logical inconsistent coefficients are highlighted in bold.

AIC indicates Akaike information criterion; BIC, Bayesian information criterion; N, number of observations.

Appendix 26: Sensitivity analysis of the DCE data

Model 5:

Main effects

| | b (R.SE) | p |
|------------------|-------------------------|-------|
| optimistic1 | -1.480*** (0.276) | 0.000 |
| optimistic2 | -0.990*** (0.226) | 0.000 |
| optimistic3 | -0.229 (0.172) | 0.182 |
| optimistic4 | -0.062 (0.099) | 0.533 |
| useful1 | -1.423*** (0.268) | 0.000 |
| useful2 | -1.054*** (0.209) | 0.000 |
| useful3 | -0.411*** (0.159) | 0.010 |
| useful4 | -0.174* (0.091) | 0.056 |
| relaxed1 | -1.347*** (0.266) | 0.000 |
| relaxed2 | -0.794*** (0.201) | 0.000 |
| relaxed3 | -0.265* (0.154) | 0.084 |
| relaxed4 | 0.019 (0.094) | 0.843 |
| dealingproblems1 | -1.368*** (0.281) | 0.000 |
| dealingproblems2 | -0.948*** (0.212) | 0.000 |
| dealingproblems3 | -0.459*** | 0.006 |

| | | |
|------------------|--------------|-------|
| | (0.166) | |
| dealingproblems4 | -0.185* | 0.052 |
| | (0.095) | |
| thinkingclearly1 | -1.216*** | 0.000 |
| | (0.253) | |
| thinkingclearly2 | -0.815*** | 0.000 |
| | (0.196) | |
| thinkingclearly3 | -0.128 | 0.378 |
| | (0.145) | |
| thinkingclearly4 | 0.117 | 0.236 |
| | (0.099) | |
| closetopeople1 | -2.229*** | 0.000 |
| | (0.275) | |
| closetopeople2 | -1.525*** | 0.000 |
| | (0.212) | |
| closetopeople3 | -0.629*** | 0.000 |
| | (0.150) | |
| closetopeople4 | -0.230** | 0.018 |
| | (0.097) | |
| makeupownmind1 | -1.213*** | 0.000 |
| | (0.274) | |
| makeupownmind2 | -0.575*** | 0.005 |
| | (0.205) | |
| makeupownmind3 | -0.452*** | 0.003 |
| | (0.154) | |
| makeupownmind4 | -0.041 | 0.657 |
| | (0.092) | |
| constant_a | -0.105** | 0.023 |
| | (0.046) | |

AIC 2787.113

BIC 2972.484

N 4412.000

Number of statistically significant main effects parameters at 5% level 19

Number of statistically 22
significant main effects
parameters at 10% level

Notes: *** significant at 1%, **significant at 5%, * significant at 10%

Robust standard errors are presented in parentheses.

Potentially logical inconsistent coefficients are highlighted in bold.

AIC indicates Akaike information criterion; BIC, Bayesian information criterion; N, number of observations.

References

- Adler, A. & Seligman, M. E. (2016) Using wellbeing for public policy: Theory, measurement, and recommendations. *International Journal of Wellbeing*, 6 (1):
- Al-Janabi, H., Flynn, T. N. & Coast, J. (2012) Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. *Quality of Life Research*, 21 (1): 167-176.
- Al-Janabi, H., Keeley, T., Mitchell, P. & Coast, J. (2013) Can capabilities be self-reported? A think aloud study. *Social Science & Medicine*, 87 116-122.
- Al-Janabi, H., McLoughlin, C., Oyebode, J., Efstathiou, N. & Calvert, M. (2019) Six mechanisms behind carer wellbeing effects: a qualitative study of healthcare delivery. *Social Science & Medicine*, 235 112382.
- Al Shabasy, S. A., Abbassi, M. M., Finch, A. P., Baines, D. & Farid, S. F. (2021) The EQ-5D-5L Valuation Study in Egypt. *Pharmacoeconomics*, 39 (5): 549-561.
- Alava, M. H., Pudney, S. & Wailoo, A. (2020) The EQ-5D-5L value set for England: findings of a quality assurance program. *Value in Health*, 23 (5): 642-648.
- Anagnostopoulos, F., Yfantopoulos, J., Moustaki, I. & Niakas, D. (2013) Psychometric and factor analytic evaluation of the 15D health-related quality of life instrument: the case of Greece. *Quality of Life Research*, 22 (8): 1973-1986.
- Anderson, N. H. & Zalinski, J. (1988) Functional measurement approach to self-estimation in multiattribute evaluation. *Journal of Behavioral Decision Making*, 1 (4): 191-221.
- Andrade, L. F., Ludwig, K., Goni, J. M. R., Oppe, M. & de Pouvourville, G. (2020) A French value set for the EQ-5D-5L. *Pharmacoeconomics*, 1-13.
- Andrich, D. (1981) PROBABILISTIC MODELS FOR SOME INTELLIGENCE AND ATTAINMENT TESTS (EXPANDED EDITION) - RASCH, G. *Applied Psychological Measurement*, 5 (4): 545-550.
- Anon (2014) *Assessment of Quality of Life (AQoL)*. [online] Available from: <https://www.aqol.com.au/index.php/what-is-aqol> (Accessed 5 November 2018).
- Anthony, R., Moore, G., Page, N., Hewitt, G., Murphy, S. & Melendez-Torres, G. (2021) Measurement invariance of the short Warwick-Edinburgh Mental Wellbeing Scale and latent mean differences (SWEMWBS) in young people by current care status. *Quality of Life Research*, 1-9.
- Arnold, D., Girling, A., Stevens, A. & Lilford, R. (2009) Comparison of direct and indirect methods of estimating health state utilities for resource allocation: review and empirical analysis. *British Medical Journal*, 339 8.

Attema, A. E., Edelaar-Peeters, Y., Versteegh, M. M. & Stolk, E. A. (2013) Time trade-off: one methodology, different methods. *The European Journal of Health Economics*, 14 (1): 53-64.

Augustovski, F., Belizán, M., Gibbons, L., Reyes, N., Stolk, E., Craig, B. M. & Tejada, R. A. (2020) Peruvian valuation of the EQ-5D-5L: a direct comparison of time trade-off and discrete choice experiments. *Value in Health*, 23 (7): 880-888.

Augustovski, F., Rey-Ares, L., Irazola, V., Garay, O. U., Gianneo, O., Fernandez, G., Morales, M., Gibbons, L. & Ramos-Goni, J. M. (2016) An EQ-5D-5L value set based on Uruguayan population preferences. *Quality of Life Research*, 25 (2): 323-333.

Augustovski, F., Rey-Ares, L., Irazola, V., Oppe, M. & Devlin, N. J. (2013) Lead versus lag-time trade-off variants: does it make any difference? *European Journal of Health Economics*, 14 S25-S31.

Bahrampour, M., Byrnes, J., Norman, R., Scuffham, P. A. & Downes, M. (2020) Discrete choice experiments to generate utility values for multi-attribute utility instruments: a systematic review of methods. *The European Journal of Health Economics*, 21 (7): 983-992.

Bailey, C., Kinghorn, P., Orlando, R., Armour, K., Perry, R., Jones, L. & Coast, J. (2016) "The ICECAP-SCM tells you more about what I'm going through": A think-aloud study measuring quality of life among patients receiving supportive and palliative care. *Palliative Medicine*, 30 (7): 642-652.

Bansback, N., Brazier, J., Tsuchiya, A. & Anis, A. (2012) Using a discrete choice experiment to estimate health state utility values. *Journal of Health Economics*, 31 (1): 306-318.

Barry, M., van Lente, E., Molcho, M., Morgan, K., McGee, H., Conroy, R., Watson, D., Shelley, E. & Perry, I. J. P. R. (2009) *SLAN 2007: Survey of Lifestyle, Attitudes and Nutrition in Ireland Mental Health and Social Well-being Report*.

Barry, M. M. (2009) Addressing the determinants of positive mental health: concepts, evidence and practice. *International Journal of Mental Health Promotion*, 11 (3): 4-17.

Bartram, D. J., Sinclair, J. M. & Baldwin, D. S. (2013) Further validation of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) in the UK veterinary profession: Rasch analysis. *Quality of Life Research*, 22 (2): 379-391.

Bartram, D. J., Yadegarfar, G., Sinclair, J. M. A. & Baldwin, D. S. (2011) Validation of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) as an overall indicator of population mental health and well-being in the UK veterinary profession. *Veterinary Journal*, 187 (3): 397-398.

Bass, M., Dawkin, M., Muncer, S., Vigurs, S. & Bostock, J. (2016) Validation of Warwick-Edinburgh mental well-being scale (WEMWBS) in a population of people using secondary care mental health services. *Journal of Mental Health*, 25 (4): 323-329.

Bech, M., Kjaer, T. & Lauridsen, J. (2011) DOES THE NUMBER OF CHOICE SETS MATTER? RESULTS FROM A WEB SURVEY APPLYING A DISCRETE CHOICE EXPERIMENT. *Health Economics*, 20 (3): 273-286.

Berlyne, D. E. (1966) Curiosity and exploration. *Science*, 153 (3731): 25-33.

Bharmal, M. & Thomas, J. (2006) Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value in Health*, 9 (4): 262-271.

Birch, S. & Donaldson, C. (2003) Valuing the benefits and costs of health care programmes: where's the 'extra' in extra-welfarism? *Social Science & Medicine*, 56 (5): 1121-1133.

Bleichrodt, H., Pinto, J. L. & Abellan-Perpignan, J. M. (2003) A consistency test of the time trade-off. *Journal of Health Economics*, 22 (6): 1037-1052.

Blumenthal-Barby, J. S. & Krieger, H. (2015) Cognitive Biases and Heuristics in Medical Decision Making: A Critical Review Using a Systematic Search Strategy. *Medical Decision Making*, 35 (4): 539-557.

Bouckaert, N., Gerkens, S., Devriese, S. & Cleemput, I. (2021) An EQ-5D-5L value set for Belgium—How to value health-related quality of life. *Health Services Research (HSR) No.*, 342

Boyatzis, R. E. (1998) *Transforming qualitative information: Thematic analysis and code development*. sage.

Boyle, M. H., Torrance, G. W., Sinclair, J. C. & Horwood, S. P. (1983) ECONOMIC-EVALUATION OF NEONATAL INTENSIVE-CARE OF VERY-LOW-BIRTH-WEIGHT INFANTS. *New England Journal of Medicine*, 308 (22): 1330-1337.

Braun, V. & Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative research in psychology*, 3 (2): 77-101.

Brazier, J. E. (2010) Is the EQ-5D fit for purpose in mental health? *British Journal of Psychiatry*, 197 (5): 348-349.

Brazier, J. E., Deverill, M. & Green, C. (1999) A review of the use of health status measures in economic evaluation. *Journal of health services research policy*, 4 (3): 174-184.

Brazier, J. E., Fukuhara, S., Roberts, J., Kharroubi, S., Yamamoto, Y., Ikeda, S., Doherty, J. & Kurokawa, K. (2009) Estimating a preference-based index from the Japanese SF-36. *Journal of Clinical Epidemiology*, 62 (12): 1323-1331.

Brazier, J. E., Ratcliffe, J., Salomon, J. & Tsuchiya, A. (2017) *Measuring and Valuing Health Benefits for Economic Evaluation*. Second edn. United Kingdom: Oxford University Press.

Brazier, J. E. & Roberts, J. (2004) The estimation of a preference-based measure of health from the SF-12. *Medical Care*, 42 (9): 851-859.

Brazier, J. E., Roberts, J. & Deverill, M. (2002) The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, 21 (2): 271-292.

Brazier, J. E., Rowen, D., Mavranouzouli, I., Tsuchiya, A., Young, T., Yang, Y., Barkham, M. & Ibbotson, R. (2012) Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome) Introduction. *Health Technology Assessment*, 16 (32): 1-+.

Brazier, J. E., Yang, Y., Tsuchiya, A. & Rowen, D. L. (2010) A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European journal of health economics*, 11 (2): 215-225.

Brent, R. J. (2014) *Cost–Benefit Analysis and Health Care Evaluations*. Second edn. Edward Elgar

Brey, P. (2012) Well-being in philosophy, psychology, and economics. In: *The good life in a technological age*. Routledge: 33-52.

Brouwer, W. B. F., Culyer, A. J., van Exel, N. J. A. & Rutten, F. F. H. (2008) Welfarism vs. extra-welfarism. *Journal of Health Economics*, 27 (2): 325-338.

Buchanan, J. & Wordsworth, S. (2015) Welfarism Versus Extra-Welfarism: Can the Choice of Economic Evaluation Approach Impact on the Adoption Decisions Recommended by Economic Evaluation Studies? *Pharmacoeconomics*, 33 (6): 571-579.

Burchardt, T. & Hick, R. (2017) Inequality and the capability approach. *Centre for Analysis of Social Exclusion, London School of Economics, Houghton Street, London WC2A 2AE*, 17.

Burk, A. (1938) A REFORMULATION OF CERTAIN ASPECTS OF WELFARE ECONOMICS. *Quarterly Journal of Economics*, 52 310-334.

Cadman, D. & Goldsmith, C. (1986) Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. *Journal of chronic diseases*, 39 (8): 643-651.

Campbell, D., Hutchinson, W. G. & Scarpa, R. (2006) Lexicographic preferences in discrete choice experiments: Consequences on individual-specific willingness to pay estimates.

Chida, Y. & Steptoe, A. (2008) Positive psychological well-being and mortality: A quantitative review of prospective observational studies. *Psychosomatic Medicine*, 70 (7): 741-756.

Chu, F., Ohinmaa, A., Klarenbach, S., Wong, Z. W. & Veugelers, P. (2017) Serum 25-Hydroxyvitamin D Concentrations and Indicators of Mental Health: An Analysis of the Canadian Health Measures Survey. *Nutrients*, 9 (10): 8.

Chua, Y. C., Wong, H. H., Abdin, E., Vaingankar, J., Shahwan, S., Cetty, L., Yong, Y. H., Hon, C., Lee, H. & Tang, C. (2020) The Recovering Quality of Life 10-item (ReQoL-10) scale in a first-episode psychosis population: Validation and implications for patient-reported outcome measures (PROMs). *Early Intervention in Psychiatry*,

Clarke, A., Friede, T., Putz, R., Ashdown, J., Martin, S., Blake, A., Adi, Y., Parkinson, J., Flynn, P., Platt, S. & Stewart-Brown, S. (2011) Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Validated for teenage school students in England and Scotland. A mixed methods assessment. *Bmc Public Health*, 11 9.

Coast, J. (2009) Maximisation in extra-welfarism: A critique of the current position in health economics. *Social Science & Medicine*, 69 (5): 786-792.

Coast, J. (2017) *Qualitative Methods for Health Economics*. Rowman & Littlefield International Limited.

Coast, J., Flynn, T. N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J. J. & Peters, T. J. (2008a) Valuing the ICECAP capability index for older people. *Social Science & Medicine*, 67 (5): 874-882.

Coast, J., Smith, R. & Lorgelly, P. (2008b) Should the capability approach be applied in health economics? *Health Economics*, 17 (6): 667-670.

Coast, J., Smith, R. D. & Lorgelly, P. (2008c) Welfarism, extra-welfarism and capability: The spread of ideas in health economics. *Social Science & Medicine*, 67 (7): 1190-1198.

Collins, D. (2014) *Cognitive interviewing practice*. Sage.

Compton, W. C., Smith, M. L., Cornish, K. A. & Qualls, D. L. (1996) Factor structure of mental health measures. *Journal of Personality and Social Psychology*, 71 (2): 406-413.

Crawford, M. J., Robotham, D., Thana, L., Patterson, S., Weaver, T., Barber, R., Wykes, T. & Rose, D. (2011) Selecting outcome measures in mental health: the views of service users. *Journal of Mental Health*, 20 (4): 336-346.

Cruz, L. N., Camey, S. A., Hoffmann, J. F., Rowen, D., Brazier, J. E., Fleck, M. P. & Polanczyk, C. A. (2011) Estimating the SF-6D Value Set for a Population-Based Sample of Brazilians. *Value in Health*, 14 (5): S108-S114.

Culyer, A. J. (1991) *THE NORMATIVE ECONOMICS OF HEALTH-CARE FINANCE AND PROVISION*. New York: Oxford University Press.

Culyer, A. J. (1995) Need: The idea won't do—But we still need it. *Social Science & Medicine*, 40 (6): 727-730.

Culyer, A. J. (2007) *Need: An Instrumental View*. Second edn.

Culyer, A. J. (2012) *Commodities, characteristics of commodities, characteristics of people, utilities, and the quality of life*. University of York.

Daly, A., Dekker, T. & Hess, S. (2016) Dummy coding vs effects coding for categorical variables: Clarifications and extensions. *Journal of Choice Modelling*, 21 36-41.

Daniel Fujiwara, Kieran Keohane, Vicky Clayton & Hotopp, U. (2017) *Mental Health and Life Satisfaction: The Relationship between the Warwick Edinburgh Mental Wellbeing Scale and Life Satisfaction* Trust, H. A. C.

Davidson, R. J. (2004) Well-being and affective style: neural substrates and biobehavioural correlates. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 359 (1449): 1395-1411.

de Bekker-Grob, E. W., Ryan, M. & Gerard, K. (2012) Discrete choice experiments in health economics: a review of the literature. *Health economics*, 21 (2): 145-172.

Deci, E. L. (1972) Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of personality and social psychology*, 22 (1): 113.

Devlin, N. J. & Brooks, R. (2017) EQ-5D and the EuroQol Group: Past, Present and Future. *Appl Health Econ Health Policy*, 15 (2): 127-137.

Devlin, N. J., Buckingham, K., Shah, K., Tsuchiya, A., Tilling, C., Wilkinson, G. & Van Hout, B. (2013) A COMPARISON OF ALTERNATIVE VARIANTS OF THE LEAD AND LAG TIME TTO. *Health Economics*, 22 (5): 517-532.

Devlin, N. J. & Krabbe, P. F. M. (2013) The development of new research methods for the valuation of EQ-5D-5L. *European Journal of Health Economics*, 14 S1-S3.

Devlin, N. J., Shah, K. K., Feng, Y., Mulhern, B. & van Hout, B. J. H. e. (2018) Valuing health-related quality of life: An EQ-5D-5L value set for England. 27 (1): 7-22.

Devlin, N. J., Shah, K. K., Mulhern, B. J., Pantiri, K. & van Hout, B. (2019) A new method for valuing health: directly eliciting personal utility functions. *The European Journal of Health Economics*, 20 (2): 257-270.

Devlin, N. J., Tsuchiya, A., Buckingham, K. & Tilling, C. (2011) A UNIFORM TIME TRADE OFF METHOD FOR STATES BETTER AND WORSE THAN DEAD: FEASIBILITY STUDY OF THE 'LEAD TIME' APPROACH. *Health Economics*, 20 (3): 348-361.

Diana Bardsley, Lucy Dean, Isla Dougall, Qingyang Feng, Lindsay Gray, Malin Karikoski, Joe Rose, Caroline Stevens & Leyland, A. H. (2017) *The Scottish Health Survey: A National Statistics Publication for Scotland*. (Accessed 20 January 2019). Government, S.

Diener, E. (1984) Subjective well-being. *Psychological Bulletin*, 95 (3): 542.

DiMatteo, M. R., Lepper, H. S. & Croghan, T. W. (2000) Depression is a risk factor for noncompliance with medical treatment - Meta-analysis of the effects of anxiety and depression on patient adherence. *Archives of Internal Medicine*, 160 (14): 2101-2107.

Dolan, P. (1997) Modeling valuations for EuroQol health states. *Medical Care*, 35 (11): 1095-1108.

Dolan, P. & Stalmeier, P. (2003) The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics*, 22 (3): 445-458.

Easterlin, R. A. (1974) Does economic growth improve the human lot? Some empirical evidence. In: *Nations and households in economic growth*. Elsevier: 89-125.

Fanshel, S. & Bush, J. W. (1970) A health-status index and its application to health-services outcomes. *Operations research*
18 (6): 1021-1066.

Feeny, D., Huguet, N., McFarland, B. H. & Kaplan, M. S. (2009) The construct validity of the Health Utilities Index Mark 3 in assessing mental health in population health surveys. *Quality of Life Research*, 18 (4): 519-526.

Feng, Y., Devlin, N. J., Shah, K. K., Mulhern, B. & van Hout, B. (2018) New methods for modelling EQ-5D-5L value sets: An application to English data. *Health Economics*, 27 (1): 23-38.

Ferreira, L. N., Ferreira, P. L., Pereira, L. N., Brazier, J. & Rowen, D. (2010) A Portuguese Value Set for the SF-6D. *Value in Health*, 13 (5): 624-630.

Ferreira, P. L., Antunes, P., Ferreira, L. N., Pereira, L. N. & Ramos-Goñi, J. M. (2019) A hybrid modelling approach for eliciting health state preferences: the Portuguese EQ-5D-5L value set. *Quality of Life Research*, 28 (12): 3163-3175.

Ferrini, S. & Scarpa, R. (2007) Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. *Journal of Environmental Economics and Management*, 53 (3): 342-363.

Finch, A. P., Meregaglia, M., Ciani, O., Roudijk, B. & Jommi, C. (2022) An EQ-5D-5L value set for Italy using videoconferencing interviews and feasibility of a new mode of administration. *Social Science & Medicine*, 292 114519.

Finnis, J. (2011) *Natural law and natural rights*. Oxford University Press.

Fletcher, G. (2013) A Fresh Start for the Objective-List Theory of Well-Being. *Utilitas*, 25 (2): 206-220.

Fletcher, G. (2016) Objective list theories. In:

Flynn, T. N., Huynh, E., Peters, T. J., Al-Janabi, H., Clemens, S., Moody, A. & Coast, J. (2015) SCORING THE ICECAP-A CAPABILITY INSTRUMENT. ESTIMATION OF A UK GENERAL POPULATION TARIFF. *Health Economics*, 24 (3): 258-269.

Flynn, T. N., Louviere, J. J., Peters, T. J. & Coast, J. (2007) Best-worst scaling: What it can do for health care research and how to do it. *Journal of Health Economics*, 26 (1): 171-189.

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P. & Grimshaw, J. M. (2010) What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health*, 25 (10): 1229-1245.

Friedli, L. & Organization, W. H. (2009) *Mental health, resilience and inequalities*.

Furlong, M. J. (2015) Social Emotional Health Survey System. *Center for School-Based Youth Development*, 5 28-2015.

Furlong, M. J., You, S., Renshaw, T. L., Smith, D. C. & O'Malley, M. D. (2014) Preliminary Development and Validation of the Social and Emotional Health Survey for Secondary School Students. *Social Indicators Research*, 117 (3): 1011-1032.

Golicki, D., Jakubczyk, M., Graczyk, K. & Niewada, M. (2019) Valuation of EQ-5D-5L health states in Poland: the first EQ-VT-based study in Central and Eastern Europe. *Pharmacoeconomics*, 37 (9): 1165-1176.

Green, C., Brazier, J. & Deverill, M. (2000) Valuing health-related quality of life - A review of health state valuation techniques. *Pharmacoeconomics*, 17 (2): 151-165.

Greene, W. H. (2003) Chapter 21: Models for discrete choice. *Econometric Analysis*, 5th ed. Upper Saddle River: Prentice Hall,

Gutierrez-Delgado, C., Galindo-Suárez, R.-M., Cruz-Santiago, C., Shah, K., Papadimitropoulos, M., Feng, Y., Zamora, B. & Devlin, N. (2021) EQ-5D-5L health-state values for the Mexican population. *Applied health economics and health policy*, 19 (6): 905-914.

Hanemann, W. M. (1984) Welfare evaluations in contingent valuation experiments with discrete responses. *American journal of agricultural economics*, 66 (3): 332-341.

Haver, A., Akerjordet, K., Caputi, P., Furunes, T. & Magee, C. (2015) Measuring mental well-being: A validation of the Short Warwick-Edinburgh Mental Well-Being Scale in Norwegian and Swedish. *Scandinavian Journal of Public Health*, 43 (7): 721-727.

Heginbotham, C. & Newbigging, K. (2013) *Commissioning health and wellbeing*. Sage.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., van Dijk, H. K. & Rotterdam, A. E. U. (2004) *Econometric Methods with Applications in Business and Economics*. OUP Oxford.

Hernández-Alava, M., Pudney, S. & Wailoo, A. (2018) Quality review of a proposed EQ-5D-5L value set for England.

Hobbins, A., Barry, L., Kelleher, D., Shah, K., Devlin, N., Goni, J. M. R. & O'Neill, C. (2018) Utility Values for Health States in Ireland: A Value Set for the EQ-5D-5L. *Pharmacoeconomics*, 36 (11): 1345-1353.

Hoffman, S., Rueda, H. A. & Lambert, M. C. (2019) Confirmatory factor analysis of the Warwick-Edinburgh Mental Wellbeing Scale among youth in Mexico. *International Social Work*, 62 (1): 309-315.

Hooker, B. (2015) The Elements of Well-Being. *Journal of Practical Ethics*, 3 (1): 15-35.

Horsman, J., Furlong, W., Feeny, D. & Torrance, G. (2003) The Health Utilities Index (HUI®): concepts, measurement properties and applications. *Health and Quality of Life Outcomes*, 1 (1): 54.

Hunt, J. (1965) Intrinsic motivation and its role in psychological development.

Huppert, F. A. & Baylis, N. (2004) Well-being: towards an integration of psychology, neurobiology and social science. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 359 (1449): 1447-1451.

I Aniza, M. H., R Otgonbayar ,Y Munkhtuul (2008) IMPORTANCE OF ECONOMIC EVALUATION IN HEALTH CARE DECISION MAKING. *Journal of Community Health*, 14

Janssen, B. M. F., Oppe, M., Versteegh, M. M. & Stolk, E. A. (2013) Introducing the composite time trade-off: a test of feasibility and face validity. *European Journal of Health Economics*, 14 S5-S13.

Jensen, C. E., Sørensen, S. S., Gudex, C., Jensen, M. B., Pedersen, K. M. & Ehlers, L. H. (2021) The Danish EQ-5D-5L Value Set: A Hybrid Model Using cTTO and DCE Data. *Applied health economics and health policy*, 1-13.

Johansson, P.-O. (1991) *An introduction to modern welfare economics*. Cambridge University Press.

Johnson, F. R., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D. A., Bresnahan, B. W., Kanninen, B. & Bridges, J. F. (2013) Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. *Value in Health*, 16 (1): 3-13.

Johnson, R., Jenkinson, D., Stinton, C., Taylor-Phillips, S., Madan, J., Stewart-Brown, S. & Clarke, A. (2016) Where's WALY? : A proof of concept study of the 'wellbeing adjusted life year' using secondary analysis of cross-sectional survey data. *Health and Quality of Life Outcomes*, 14 9.

Jones-Lee, M., Loomes, G., O'Reilly, D. & Philips, P. (1993) The value of preventing non-fatal road injuries: findings of a willingness-to-pay national sample survey. *Transport Research Laboratory Contractor Report*, (CR 330):

Kahneman, D., Diener, E. & Schwarz, N. (1999) *Well-being: Foundations of hedonic psychology*. Russell Sage Foundation.

Kahneman, D., Slovic, S. P., Slovic, P. & Tversky, A. (1982) *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kammann, R. & Flett, R. (1983) AFFECTOMETER-2 - A SCALE TO MEASURE CURRENT LEVEL OF GENERAL HAPPINESS. *Australian Journal of Psychology*, 35 (2): 259-265.

Kaplan, R. M. & Anderson, J. P. (1996) The general health policy model: an integrated approach. *Quality of life pharmacoeconomics in clinical trials* 2302-322.

Kaplan, R. M., Anderson, J. P. & Ganiats, T. G. (1993) The quality of well-being scale: rationale for a single quality of life index. In: *Quality of life assessment: key issues in the 1990s*. Springer: 65-94.

Kaplan, R. M., Bush, J. W. & Berry, C. C. (1976) Health status: types of validity and the index of well-being. *Health services research*, 11 (4): 478.

Kaplan, R. M., Bush, J. W. & Berry, C. C. (1979) HEALTH-STATUS INDEX - CATEGORY RATING VERSUS MAGNITUDE ESTIMATION FOR MEASURING LEVELS OF WELL-BEING. *Medical Care*, 17 (5): 501-525.

Karimi, M., Brazier, J. & Basarir, H. (2016) The Capability Approach: A Critical Review of Its Application in Health Economics. *Value in Health*, 19 (6): 795-799.

Keeney, R. L. & Raiffa, H. (1993) *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

Keetharuth, A., Brazier, J., Connell, J., Carlton, J., Taylor Buck, E., Ricketts, T. & Barkham, M. (2017) Development and validation of the Recovering Quality of Life (ReQoL) outcome measures. In: *EEPRU Technical Research Report 050. Policy Research Unit in Economic Evaluation of Health and Care Interventions*. University of Sheffield and York:

Keetharuth, A. D., Brazier, J., Connell, J., Bjorner, J. B., Carlton, J., Buck, E. T., Ricketts, T., McKendrick, K., Browne, J. & Croudace, T. (2018a) Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. *The British Journal of Psychiatry*, 212 (1): 42-49.

Keetharuth, A. D., Rowen, D., Bjorner, J. B. & Brazier, J. (2021) Estimating a preference-based index for mental health from the recovering quality of life measure: Valuation of recovering quality of life utility index. *Value in Health*, 24 (2): 281-290.

Keetharuth, A. D., Taylor Buck, E., Acquadro, C., Conway, K., Connell, J., Barkham, M., Carlton, J., Ricketts, T., Barber, R. & Brazier, J. (2018b) Integrating qualitative and quantitative data in the development of outcome measures: The case of the Recovering Quality of Life (ReQoL) measures in mental health populations. *International Journal of Environmental Research and Public Health*, 15 (7): 1342.

Keyes, C. L. (2009) Brief description of the mental health continuum short form (MHC-SF).

Keyes, C. L. M. (2013) *Mental Well-Being: International Contributions to the Study of Positive Mental Health*. Springer.

Kim, S. H., Ahn, J., Ock, M., Shin, S., Park, J., Luo, N. & Jo, M. W. (2016) The EQ-5D-5L valuation study in Korea. *Quality of Life Research*, 25 (7): 1845-1852.

King, L. A. & Napa, C. K. (1998) What makes a life good? *Journal of Personality and Social Psychology*, 75 (1): 156-165.

Koushede, V., Lasgaard, M., Hinrichsen, C., Meilstrup, C., Nielsen, L., Rayce, S. B., Torres-Sahli, M., Gudmundsdottir, D. G., Stewart-Brown, S. & Santini, Z. I. (2019) Measuring mental well-being in Denmark: Validation of the original and short version of the Warwick-Edinburgh mental well-being scale (WEMWBS and SWEMWBS) and cross-cultural comparison across four European settings. *Psychiatry Research*, 271 502-509.

Krabbe, P. F. M., Devlin, N. J., Stolk, E. A., Shah, K. K., Oppe, M., van Hout, B., Quik, E. H., Pickard, A. S. & Xie, F. (2014) Multinational Evidence of the Applicability and Robustness of

Discrete Choice Modeling for Deriving EQ-5D-5L Health-State Values. *Medical Care*, 52 (11): 935-943.

Krucien, N., Watson, V. & Ryan, M. (2017) Is Best-Worst Scaling Suitable for Health State Valuation? A Comparison with Discrete Choice Experiments. *Health Economics*, 26 (12): E1-E16.

Lam, C. L. K., Brazier, J. & McGhee, S. M. (2008) Valuation of the SF-6D health states is feasible, acceptable, reliable, and valid in a Chinese population. *Value in Health*, 11 (2): 295-303.

Lamers, L., Bouwmans, C., van Straten, A., Donker, M. & Hakkaart, L. (2006) Comparison of EQ-5D and SF-6D utilities in mental health patients. *Health economics*, 15 (11): 1229-1236.

Lancaster, K. J. (1966) A new approach to consumer theory. *Journal of political economy*, 74 (2): 132-157.

Lancsar, E., Fiebig, D. G. & Hole, A. R. (2017) Discrete choice experiments: a guide to model specification, estimation and software. *Pharmacoeconomics*, 35 (7): 697-716.

Lancsar, E. & Louviere, J. (2008) Conducting discrete choice experiments to inform Healthcare decision making. *Pharmacoeconomics*, 26 (8): 661-677.

Le Galès, C., Buron, C., Costet, N., Rosman, S. & Slama, P. G. (2002) Development of a preference-weighted health status classification system in France: the Health Utilities Index 3. *Health care management science*, 5 (1): 41-51.

Le, Q. A., Doctor, J. N., Zoellner, L. A. & Feeny, N. C. (2013) Minimal clinically important differences for the EQ-5D and QWB-SA in Post-traumatic Stress Disorder (PTSD): results from a Doubly Randomized Preference Trial (DRPT). *Health and Quality of Life Outcomes*, 11 9.

Lee, C. H., Cook, S., Lee, J. S. & Han, B. (2016) Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of Z-scores. *Genomics & informatics*, 14 (4): 173.

Lenert, L. A., Cher, D. J., Goldstein, M. K., Bergen, M. R. & Garber, A. (1998) The effect of search procedures on utility elicitations. *Medical Decision Making*, 18 (1): 76-83.

Leppanen, V., Hakko, H., Sintonen, H. & Lindeman, S. (2016) Comparing Effectiveness of Treatments for Borderline Personality Disorder in Communal Mental Health Care: The Oulu BPD Study. *Community Mental Health Journal*, 52 (2): 216-227.

Lin, H.-W., Li, C.-I., Lin, F.-J., Chang, J.-Y., Gau, C.-S., Luo, N., Pickard, A. S., Ramos Goñi, J. M., Tang, C.-H. & Hsu, C.-N. (2018) Valuation of the EQ-5D-5L in Taiwan. *PLoS One*, 13 (12): e0209344.

Lizzie Trotter, Mary-Kathryn Rallings Adams, Daniel Fujiwara, Kieran Keohane & Clayton, V. (2017) *Valuing improvements in mental health: Applying the wellbeing valuation method to WEMWBS*. Trust, H. A. C.

Long, J. S. & Long, J. S. (1997) *Regression models for categorical and limited dependent variables*. Sage.

Louviere, J. J., Hensher, D. A., Swait, J. D. & Adamowicz, W. (2000) *Stated Choice Methods: Analysis and Applications*. Cambridge University Press.

Louviere, J. J. & Woodworth, G. (1983) Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *Journal of marketing research*, 20 (4): 350-367.

Ludwig, K., von der Schulenburg, J. M. G. & Greiner, W. (2018) German Value Set for the EQ-5D-5L. *Pharmacoeconomics*, 36 (6): 663-674.

Lugnér, A. K. & Krabbe, P. F. (2020) An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. *Expert review of pharmacoeconomics & outcomes research*, 20 (4): 331-342.

Luo, N., Liu, G., Li, M. H., Guan, H. J., Jin, X. J. & Rand-Hendriksen, K. (2017) Estimating an EQ-5D-5L Value Set for China. *Value in Health*, 20 (4): 662-669.

Luo, N., Seng, B. K., Thumboo, J., Feeny, D. & Li, S. C. (2006) A study of the construct validity of the health utilities index mark 3 (HUI3) in patients with schizophrenia. *Quality of Life Research*, 15 (5): 889-898.

Lyubomirsky, S., King, L. & Diener, E. (2005) The benefits of frequent positive affect: Does happiness lead to success? *Psychological bulletin*, 131 (6): 803.

Maheswaran, H., Weich, S., Powell, J. & Stewart-Brown, S. (2012) Evaluating the responsiveness of the Warwick Edinburgh Mental Well-Being Scale (WEMWBS): Group and individual level analysis. *Health and Quality of Life Outcomes*, 10 8.

Mai, V. Q., Sun, S., Van Minh, H., Luo, N., Giang, K. B., Lindholm, L. & Sahlen, K. G. (2020) An EQ-5D-5L value set for Vietnam. *Quality of Life Research*, 29 (7): 1923-1933.

Marley, A. A. J. & Louviere, J. J. (2005) Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, 49 (6): 464-480.

Mavranouzouli, I., Brazier, J. E., Rowen, D. & Barkham, M. (2013) Estimating a preference-based index from the Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM) valuation of CORE-6D. *Medical Decision Making*, 33 (3): 381-395.

- McFadden, D. (1973) Conditional logit analysis of qualitative choice behavior.
- McIntosh, E. & Louviere, J. (2002) Separating weight and scale value: an exploration of best-attribute scaling in health economics. *Health Economics Study Group. Odense, Denmark*,
- McMahan, E. A. & Estes, D. (2011) Hedonic Versus Eudaimonic Conceptions of Well-being: Evidence of Differential Associations With Self-reported Well-being. *Social Indicators Research*, 103 (1): 93-108.
- Mehrez, A. & Gafni, A. (1991) THE HEALTHY-YEARS EQUIVALENTS - HOW TO MEASURE THEM USING THE STANDARD GAMBLE APPROACH. *Medical Decision Making*, 11 (2): 140-146.
- Mendez, I., Perpignan, J. M. A., Martinez, F. I. S. & Perez, J. E. M. (2011) Inverse probability weighted estimation of social tariffs: An illustration using the SF-6D value sets. *Journal of Health Economics*, 30 (6): 1280-1292.
- Michael F. Drummond, Mark J. Sculpher, Karl Claxton, Greg L. Stoddart & Torrance, G. W. (2015) *Methods for the Economic Evaluation of Health Care Programmes*.
- Morgenstern, O. & Von Neumann, J. (1953) *Theory of games and economic behavior*. Princeton university press.
- Mukuria, C., Peasgood, T. & Brazier, J. (2021) Applying EuroQol Portable Valuation Technology to the EQ Health and Wellbeing Short (EQHWB-S): a pilot study. *School of Health and Related Research, University of Sheffield Discussion Paper Series*,
- Mulhern, B., Norman, R., Street, D. J. & Viney, R. (2019) One method, many methodological choices: a structured review of discrete-choice experiments for health state valuation. *Pharmacoeconomics*, 37 (1): 29-43.
- Murphy, M. C. (2001) *Natural law and practical rationality*. Cambridge University Press.
- Murtagh, F. E. M., Addington-Hall, J. M. & Higginson, I. J. (2007) The value of cognitive interviewing techniques in palliative care research. *Palliative Medicine*, 21 (2): 87-93.
- Netten, A., Burge, P., Malley, J., Potoglou, D., Towers, A. M., Brazier, J., Flynn, T., Forder, J. & Wall, B. (2012) Outcomes of social care for adults: developing a preference-weighted measure. *Health Technology Assessment*, 16 (16): 1-+.
- Ng Fat, L., Scholes, S., Boniface, S., Mindell, J. & Stewart-Brown, S. (2017) Evaluating and establishing national norms for mental wellbeing using the short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS): findings from the Health Survey for England. *Quality of Life Research*, 26 (5): 1129-1144.

Ng, S. S., Lo, A. W., Leung, T. K., Chan, F. S., Wong, A. T., Lam, R. W. & Tsang, D. K. (2014) Translation and validation of the Chinese version of the short Warwick-Edinburgh Mental Well-being Scale for patients with mental illness in Hong Kong. *East Asian Archives of Psychiatry*, 24 (1): 3-9.

Nord, E. (1995) THE PERSON-TRADE-OFF APPROACH TO VALUING HEALTH-CARE PROGRAMS. *Medical Decision Making*, 15 (3): 201-208.

Norman, R., Viney, R., Brazier, J., Burgess, L., Cronin, P., King, M., Ratcliffe, J. & Street, D. (2014) Valuing SF-6D Health States Using a Discrete Choice Experiment. *Medical Decision Making*, 34 (6): 773-786.

O'Sullivan, E. & Schofield, S. (2018) Cognitive bias in clinical medicine. *JR Coll Physicians Edinb*, 48 (3): 225-232.

Office for National Statistics *2011 Census: Key Statistics and Quick Statistics for Local Authorities in the United Kingdom*. [online] Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/keystatisticsandquickstatisticsforlocalauthoritiesintheunitedkingdom/2013-10-11> (Accessed 12/10/2021).

Office for National Statistics *Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2020*. [online] Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2020> (Accessed 12/10/2021).

Oppe, M., Devlin, N. J., van Hout, B., Krabbe, P. F. M. & de Charro, F. (2014) A Program of Methodological Research to Arrive at the New International EQ-5D-5L Valuation Protocol. *Value in Health*, 17 (4): 445-453.

Oppe, M., Rand-Hendriksen, K., Shah, K., Ramos-Goni, J. M. & Luo, N. (2016) EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. *Pharmacoeconomics*, 34 (10): 993-1004.

Oppe, M. & Van Hout, B. (2010) The optimal hybrid: experimental design and modeling of a combination of TTO and DCE.

Oppe, M. & Van Hout, B. (2017) The “power” of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. *EuroQol Working Paper Series*, 17003

Papadimitropoulos, E. A., Elbarazi, I., Blair, I., Katsaiti, M.-S., Shah, K. K. & Devlin, N. J. (2015) An investigation of the feasibility and cultural appropriateness of stated preference methods to generate health state values in the United Arab Emirates. *Value in health regional issues*, 7 34-41.

Parducci, A. (1995) *Happiness, pleasure, and judgment: the contextual theory and its applications*. Lawrence Erlbaum Associates, Inc.

Parfit, D. (1984) *Reasons and persons*. OUP Oxford.

Parkinson, J. (2007) Establishing a core set of national, sustainable mental health indicators for adults in Scotland: Rationale paper. *Glasgow: NHS Health Scotland*,

Pattanaphesaj, J., Thavorncharoensap, M., Ramos-Goni, J. M., Tongsir, S., Ingsrisawang, L. & Teerawattananon, Y. (2018) The EQ-5D-5L Valuation study in Thailand. *Expert Review of Pharmacoeconomics & Outcomes Research*, 18 (5): 551-558.

Perpinan, J. M. A., Martinez, F. I. S., Perez, J. E. M. & Mendez, I. (2012) LOWERING THE 'FLOOR' OF THE SF-6D SCORING ALGORITHM USING A LOTTERY EQUIVALENT METHOD. *Health Economics*, 21 (11): 1271-1285.

Pickard, A. S., Law, E. H., Jiang, R., Pullenayegum, E., Shaw, J. W., Xie, F., Oppe, M., Boye, K. S., Chapman, R. H. & Gong, C. L. (2019) United States valuation of EQ-5D-5L health states using an international protocol. *Value in Health*, 22 (8): 931-941.

Pliskin, J. S., Shepard, D. S. & Weinstein, M. C. (1980) UTILITY-FUNCTIONS FOR LIFE YEARS AND HEALTH-STATUS. *Operations Research*, 28 (1): 206-224.

Potoglou, D., Burge, P., Flynn, T., Netten, A., Malley, J., Forder, J. & Brazier, J. E. (2011) Best-worst scaling vs. discrete choice experiments: An empirical comparison using social care data. *Social Science & Medicine*, 72 (10): 1717-1727.

Powell, J., Hamborg, T., Stallard, N., Burls, A., McSorley, J., Bennett, K., Griffiths, K. M. & Christensen, H. (2013) Effectiveness of a web-based cognitive-behavioral tool to improve mental well-being in the general population: randomized controlled trial. *Journal of medical Internet research*, 15 (1): e2240.

Prades, J. L. P. (1997) Is the person trade-off a valid method for allocating health care resources? *Health Economics*, 6 (1): 71-81.

Pressman, S. D. & Cohen, S. (2005) Does positive affect influence health? *Psychological Bulletin*, 131 (6): 925-971.

Purba, F. D., Hunfeld, J. A. M., Iskandarsyah, A., Fitriana, T. S., Sadarjoen, S. S., Ramos-Goni, J. M., Passchier, J. & Busschbach, J. J. V. (2017) The Indonesian EQ-5D-5L Value Set. *Pharmacoeconomics*, 35 (11): 1153-1165.

Pyne, J. M., Sieber, W. J., David, K., Kaplan, R. M., Rapaport, M. H. & Williams, D. K. (2003) Use of the quality of well-being self-administered version (QWB-SA) in assessing health-related quality of life in depressed patients. *Journal of Affective Disorders*

76 (1-3): 237-247.

Ramos-Goñi, J., Craig, A., Oppe, M. & Van Hout, B. (2016) Combining continuous and dichotomous responses in a hybrid model. *Improving the Valuation of the EQ-5D-5L by Introducing Quality Control and Integrating TTO and DCE*, 133.

Ramos-Goni, J. M., Craig, B. M., Oppe, M., Ramallo-Farina, Y., Pinto-Prades, J. L., Luo, N. & Rivero-Arias, O. (2018) Handling Data Quality Issues to Estimate the Spanish EQ-5D-5L Value Set Using a Hybrid Interval Regression Approach. *Value in Health*, 21 (5): 596-604.

Ramos-Goni, J. M., Oppe, M., Slaap, B., Busschbach, J. J. V. & Stolk, E. (2017a) Quality Control Process for EQ-5D-5L Valuation Studies. *Value in Health*, 20 (3): 466-473.

Ramos-Goni, J. M., Pinto-Prades, J. L., Oppe, M., Cabases, J. M., Serrano-Aguilar, P. & Rivero-Arias, O. (2017b) Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. *Medical Care*, 55 (7): E51-E58.

Reed, W. W., Herbers, J. E. & Noel, G. L. (1993) CHOLESTEROL-LOWERING THERAPY - WHAT PATIENTS EXPECT IN RETURN. *Journal of General Internal Medicine*, 8 (11): 591-596.

Rencz, F., Brodszky, V., Gulácsi, L., Golicki, D., Ruzsa, G., Pickard, A. S., Law, E. H. & Péntek, M. (2020) Parallel valuation of the EQ-5D-3L and EQ-5D-5L by time trade-off in Hungary. *Value in Health*, 23 (9): 1235-1245.

Rice, C. M. (2013) Defending the Objective List Theory of Well-Being. *Ratio*, 26 (2): 196-211.

Richardson, J. (1994) COST-UTILITY ANALYSIS - WHAT SHOULD BE MEASURED. *Social Science & Medicine*, 39 (1): 7-21.

Richardson, J., Elsworth, G., Iezzi, A., Khan, M. A., Mihalopoulos, C., Schweitzer, I. & Herrman, H. (2011a) Increasing the sensitivity of the AQoL inventory for evaluation of interventions affecting mental health. *Research Paper*, 61

Richardson, J., McKie, J. & Bariola, E. (2011b) *Review and critique of health related multi attribute utility instruments*. Monash University, Business and Economics, Centre for Health Economics Melbourne.

Richardson, J., Sinha, K., Iezzi, A. & Khan, M. A. (2014) Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. *Quality of Life Research*, 23 (8): 2395-2404.

Robeyns, I. (2005) The capability approach: a theoretical survey. *Journal of human development* 6(1): 93-117.

Robin W. Boadway & Bruce, N. (1984) *Welfare Economics*. England: Basil Blackwell Publisher Limited.

Robinson, A., Dolan, P. & Williams, A. (1997) Valuing health status using VAS and TTO: What lies behind the numbers? *Social Science & Medicine*, 45 (8): 1289-1297.

Robinson, A. & Spencer, A. (2006) Exploring challenges to TTO utilities: valuing states worse than dead. *Health Economics*, 15 (4): 393-402.

Rogers, K. D., Dodds, C., Campbell, M. & Young, A. (2018) The validation of the Short Warwick-Edinburgh Mental Well-Being Scale (SWEMWBS) with deaf British sign language users in the UK. *Health and Quality of Life Outcomes*, 16 (1): 145.

Rose, J. M. & Bliemer, M. C. (2004) The design of stated choice experiments: The state of practice and future challenges.

Rose, T., Joe, S., Williams, A., Harris, R., Betz, G. & Stewart-Brown, S. (2017) Measuring mental wellbeing among adolescents: a systematic review of instruments. *Journal of Child Family Studies* 26 (9): 2349-2362.

Rowen, D., Brazier, J. & Van Hout, B. (2015) A Comparison of Methods for Converting DCE Values onto the Full Health-Dead QALY Scale. *Medical Decision Making*, 35 (3): 328-340.

Ryan, M. (2004) Discrete choice experiments in health care.

Ryan, M. & Gerard, K. (2003) Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied health economics and health policy*, 2 (1): 55-64.

Ryan, M., Watson, V. & Entwistle, V. (2009) Rationalising the 'irrational': a think aloud study of discrete choice experiment responses. *Health economics*, 18 (3): 321-336.

Ryan, R. M. & Deci, E. L. (2001) On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual review of psychology*, 52 (1): 141-166.

Ryff, C. D. (1995) PSYCHOLOGICAL WELL-BEING IN ADULT LIFE. *Current Directions in Psychological Science*, 4 (4): 99-104.

Saarni, S. I., Viertio, S., Perala, J., Koskinen, S., Lonnqvist, J. & Suvisaari, J. (2010) Quality of life of people with schizophrenia, bipolar disorder and other psychotic disorders. *British Journal of Psychiatry*, 197 (5): 386-394.

Samuelson, P. A. (1947) *Foundations of Economic Analysis*. Harvard University Press.

Scitovsky, T. (1976) *An Inquiry into human satisfaction and consumer dissatisfaction*.

Scottish Government (2018) *National indicator performance*. [online] Available from: <https://nationalperformance.gov.scot/measuring-progress/national-indicator-performance> (Accessed 3 January).

Seiber, W. J., Groessl, E. J., David, K. M., Ganiats, T. G. & Kaplan, R. M. (2008) Quality of well being self-administered (QWB-SA) scale. *Health Services Research Center, University of California, San Diego*,

Seixas, B. V. (2017) Welfarism and extra-welfarism: a critical overview. *Cadernos De Saude Publica*, 33 (8): 9.

Sen, A. (1979) Personal Utilities and Public Judgements: Or What's Wrong With Welfare Economics. *The Economic Journal*, 89 (355): 537-558.

Sen, A. (1980) Equality of What? In: McMurrin, S., ed. *Tanner Lectures on Human Values, Volume 1*. Cambridge: Cambridge University Press:

Sen, A. (1982) *Choice, Welfare and Measurement*. Cambridge: Harvard University Press.

Sen, A. (1993) Capability and well-being. In: Oxford: Clarendon Press:

Shafie, A. A., Thakumar, A. V., Lim, C. J., Luo, N., Rand-Hendriksen, K. & Yusof, F. A. M. (2018) EQ-5D-5L Valuation for the Malaysian Population. *Pharmacoeconomics*, 1-11.

Shah, K., Rand-Hendriksen, K., Ramos-Goni, J., Prause, A. & Stolk, E. (2014) Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme.

Shah, K. K., Mulhern, B., Longworth, L. & Janssen, M. F. (2017a) Views of the UK General Public on Important Aspects of Health Not Captured by EQ-5D. *Patient-Patient Centered Outcomes Research*, 10 (6): 701-709.

Shah, N., Cader, M., Andrews, B., McCabe, R. & Stewart-Brown, S. L. (2021) Short Warwick-Edinburgh Mental Well-being Scale (SWEMWBS): performance in a clinical sample in relation to PHQ-9 and GAD-7. *Health and Quality of Life Outcomes*, 19 (1): 1-9.

Shah, N., Steiner, D., Petrou, S., Johnson, R. & Stewart Brown, S. (2017b) Exploring the impact of the Warwick-Edinburgh Mental Well-being scales on public health research and practice (in press 2018). *Health Services Research and Policy*,

Shah, N. & Stewart-Brown, S. (2017) The Warwick-Edinburgh Mental Wellbeing Scale: role and impact on public health policy and practice. *European Journal of Public Health*, 27 484-484.

Shiroiwa, T., Ikeda, S., Noto, S., Igarashi, A., Fukuda, T., Saito, S. & Shimozuma, K. (2016) Comparison of Value Set Based on DCE and/or TTO Data: Scoring for EQ-5D-5L Health States in Japan. *Value in Health*, 19 (5): 648-654.

Sintonen, H. (1995) The 15D-measure of health-related quality of life. II. Feasibility, reliability and validity of its valuation system. *National Centre for Health Program Evaluation Working Paper 42*, Melbourne

Sintonen, H. (2001) The 15D instrument of health-related quality of life: properties and applications. *Annals of Medicine*, 33 (5): 328-336.

Spencer, A. (2003) The TTO method and procedural invariance. *Health Economics*, 12 (8): 655-668.

Steptoe, A., Wardle, J. & Marmot, M. (2005) Positive affect and health-related neuroendocrine, cardiovascular, and inflammatory processes. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (18): 6508-6512.

Stewart-Brown, S. (2021) *15 years on: Insights and reflections on the Warwick-Edinburgh Mental Wellbeing Scales (WEMWBS)*. What Works Centre for Wellbeing.

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J. & Weich, S. (2009) Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey. *Health and Quality of Life Outcomes*, 7 8.

Stolk, E., Ludwig, K., Rand, K., van Hout, B. & Ramos-Goni, J. M. (2019) Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol. *Value in Health*, 22 (1): 23-30.

Stolk, E. A., Oppe, M., Scalone, L. & Krabbe, P. F. M. (2010) Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. *Value in Health*, 13 (8): 1005-1013.

Sullivan, T., Hansen, P., Ombler, F., Derrett, S. & Devlin, N. (2020) A new tool for creating personal and social EQ-5D-5L value sets, including valuing 'dead'. *Social Science & Medicine*, 246 112707.

Taggart, F., Friede, T., Weich, S., Clarke, A., Johnson, M. & Stewart-Brown, S. (2013) Cross cultural evaluation of the Warwick-Edinburgh mental well-being scale (WEMWBS) -a mixed methods study. *Health and Quality of Life Outcomes*, 11 12.

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J. & Stewart-Brown, S. (2007a) The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. *Health and Quality of Life Outcomes*, 5 13.

Tennant, R., Joseph, S. & Stewart-Brown, S. (2007b) The Affectometer 2: a measure of positive mental health in UK populations. *Quality of Life Research*, 16 (4): 687-695.

Tesio, L. (2003) Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine*, 35 (3): 105-115.

The National Institute for Health and Care Excellence (2013) Guide to the methods of technology appraisal 2013.

Tilling, C., Devlin, N., Tsuchiya, A. & Buckingham, K. (2008) Protocols for TTO Valuations of Health States Worse than Dead: A literature review and framework for systematic analysis.

Tolley, K. (2009) What are health utilities. *London: Hayward Medical Communications*,

Topp, C. W., Ostergaard, S. D., Sondergaard, S. & Bech, P. (2015) The WHO-5 Well-Being Index: A Systematic Review of the Literature. *Psychotherapy and Psychosomatics*, 84 (3): 10.

Torrance, G. W. (1976) Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-economic planning sciences*, 10 (3): 129-136.

Torrance, G. W. (1986) MEASUREMENT OF HEALTH STATE UTILITIES FOR ECONOMIC APPRAISAL - A REVIEW. *Journal of Health Economics*, 5 (1): 1-30.

Torrance, G. W., Boyle, M. H. & Horwood, S. P. (1982) APPLICATION OF MULTI-ATTRIBUTE UTILITY-THEORY TO MEASURE SOCIAL PREFERENCES FOR HEALTH STATES. *Operations Research*, 30 (6): 1043-1069.

Tversky, A. & Kahneman, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5 (2): 207-232.

Vaingankar, J. A., Abidin, E., Chong, S. A., Sambasivam, R., Seow, E., Jeyagurunathan, A., Picco, L., Stewart-Brown, S. & Subramaniam, M. (2017) Psychometric properties of the short Warwick Edinburgh mental well-being scale (SWEMWBS) in service users with schizophrenia, depression and anxiety spectrum disorders. *Health and Quality of Life Outcomes*, 15 (1): 153.

van Hout, B., Janssen, M. F., Feng, Y. S., Kohlmann, T., Busschbach, J., Golicki, D., Lloyd, A., Scalone, L., Kind, P. & Pickard, A. S. (2012) Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets. *Value in Health*, 15 (5): 708-715.

Vanderdonk, J., Levendag, P. C., Kuijpers, A. J., Roest, F. H. J., Habbema, J. D. F., Meeuwis, C. A. & Schmitz, P. I. M. (1995) PATIENT PARTICIPATION IN CLINICAL DECISION-MAKING FOR TREATMENT OF T3 LARYNGEAL-CANCER - A COMPARISON OF STATE AND PROCESS UTILITIES. *Journal of Clinical Oncology*, 13 (9): 2369-2378.

Veldwijk, J., Lambooi, M. S., de Bekker-Grob, E. W., Smit, H. A. & de Wit, G. A. (2014) The Effect of Including an Opt-Out Option in Discrete Choice Experiments. *PLoS One*, 9 (11): 9.

Versteegh, M. M., Attema, A. E., Oppe, M., Devlin, N. J. & Stolk, E. A. (2013) Time to tweak the TTO: results from a comparison of alternative specifications of the TTO. *European Journal of Health Economics*, 14 S43-S51.

Versteegh, M. M., Vermeulen, K. M., Evers, S., de Wit, G. A., Prenger, R. & Stolk, E. A. (2016) Dutch Tariff for the Five-Level Version of EQ-5D. *Value in Health*, 19 (4): 343-352.

Waqas, A., Ahmad, W., Haddad, M., Taggart, F. M., Muhammad, Z., Bukhari, M. H., Sami, S. A., Batool, S. M., Najeeb, F., Hanif, A., Rizvi, Z. A. & Ejaz, S. (2015) Measuring the well-being of health care professionals in the Punjab: a psychometric evaluation of the Warwick-Edinburgh Mental Well-being Scale in a Pakistani population. *PeerJ*, 3 15.

Welie, A. G., Gebretekle, G. B., Stolk, E., Mukuria, C., Krahn, M. D., Enquoselassie, F. & Fenta, T. G. (2020) Valuing health state: an EQ-5D-5L value set for Ethiopians. *Value in health regional issues*, 22 7-14.

White, R. W. (1959) Motivation reconsidered: The concept of competence. *Psychological review*, 66 (5): 297.

Whitehead, S. J. & Ali, S. (2010) Health outcomes in economic evaluation: the QALY and utilities. *British Medical Bulletin*, 96 (1): 5-21.

Whitty, J. A., Ratcliffe, J., Chen, G. & Scuffham, P. A. (2014) Australian Public Preferences for the Funding of New Health Technologies: A Comparison of Discrete Choice and Profile Case Best-Worst Scaling Methods. *Medical Decision Making*, 34 (5): 638-654.

Williams, A. (2005) EQ-5D concepts and methods: a developmental history.

Willis, G. B. (1994) *Cognitive interviewing and questionnaire design: a training manual*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Willis, G. B. (2004) *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.

Wong, E. L. Y., Ramos-Goni, J. M., Cheung, A. W. L., Wong, A. Y. K. & Rivero-Arias, O. (2018) Assessing the Use of a Feedback Module to Model EQ-5D-5L Health States Values in Hong Kong. *Patient-Patient Centered Outcomes Research*, 11 (2): 235-247.

Wooldridge, J. M. (2015) *Introductory econometrics: A modern approach*. Cengage learning.

Xie, F., Pullenayegum, E., Gaebel, K., Bansback, N., Bryan, S., Ohinmaa, A., Poissant, L. & Johnson, J. A. (2016) A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. *Medical Care*, 54 (1): 98-105.

Yang, F., Katumba, K. R., Roudijk, B., Yang, Z., Reville, P., Griffin, S., Ochanda, P. N., Lamorde, M., Greco, G. & Seeley, J. (2021) Developing the EQ-5D-5L Value Set for Uganda Using the 'Lite' Protocol. *Pharmacoeconomics*, 1-13.

Yang, Z., Feng, Z., Busschbach, J., Stolk, E. & Luo, N. (2019) How prevalent are implausible EQ-5D-5L health states and how do they affect valuation? A study combining quantitative and qualitative evidence. *Value in Health*, 22 (7): 829-836.

Yang, Z. H., van Busschbach, J., Timman, R., Janssen, M. F. & Luo, N. (2017) Logical inconsistencies in time trade-off valuation of EQ-5D-5L health states: Whose fault is it? *PLoS One*, 12 (9): 10.

Young, T. A., Rowen, D., Norquist, J. & Brazier, J. E. (2010) Developing preference-based health measures: using Rasch analysis to generate health state values. *Quality of Life Research*, 19 (6): 907-917.