Broadening the Measurement and Valuation of Health and Quality of Life

Brendan Mulhern

Doctor of Philosophy

Centre for Health Economics Research and Evaluation

Faculty of Business

University of Technology Sydney

First submitted in November 2019

Resubmitted in July 2020

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, BRENDAN MULHERN declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Business at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 03/07/2020

ACKNOWLEDGEMENTS

I would like to thank many people who have helped me complete this thesis over the last three or so years. Thanks to my supervisors, Prof's Rosalie Viney and Deborah Street for all the support, and for the insightful comments that have immeasurably improved this work. In particular I thank Rosalie for giving me the chance to move to CHERE in 2015 and encouraging me to (eventually) start my PhD. Particular thanks to Debbie for making me think outside of the box when it came to DCE methods, challenging my analysis skills, and for the long discussions in both of our offices about my thesis and beyond. Madeleine King assessed each stage of my thesis, and her comments were incredibly useful.

I would also like to acknowledge my career mentors Prof. John Brazier and Prof. Aki Tsuchiya for the ongoing research collaborations and support. Also, thanks to my collaborators on the projects I have included in my thesis, and also many other health economists around the world that have directly or indirectly helped me over the line. In particular, thanks go to Richard Norman and Richard De Abreu Lourenco for the advice about academic and PhD life in Australia, and for contributing to the papers published in this thesis. Juliette Malley and Mark Oppe also helped to improve the studies included. The EuroQol Group have also been incredibly supportive of the sort of empirical work reported here, and thanks to the group for inspirational discussions at the various annual meetings. Work in this PhD was funded by the EuroQol Group, and I thank them for financial support as well Nick Bansback and Koonal Shah have also encouraged me through to the end, and I value their support.

I would also like to thank everyone at CHERE for their support, and giving me space with my office door closed (and taking work off my hands) when it came towards the end! In particular I thank Phil Haywood, Tom Longden and Kathleen Manipis for the support, and distractions when required.

Since moving to Sydney, and during my PhD, I have been lucky to meet my now fiancée Dr. Emily Trimmer. I thank her for her support in all aspects of our lives.

I also thank my parents in England and my future in laws for their support. Also, to friends outside academia with no interest in health economics that allowed for an escape from this thesis when needed.

THESIS FORMAT

This is a conventional thesis including seven chapters of content, references and further appendices. Two of the chapters (a literature review and empirical study) three empirical studies have resulted in published journal articles as described below.

PUBLISHED PAPERS RESULTING FROM THIS THESIS

1. **Mulhern B**, Norman R, DeAbreu Lourenco R, Street D, Malley J, Viney R. Investigating the relative value of health and social care related quality of life using a Discrete Choice Experiment. Social Science and Medicine. 2019; 233: 28-37.

   Author contribution for this paper: Brendan Mulhern led the conceptualisation of the study, developed the study design, constructed the experiment, and led the data analysis and interpretation and paper writing. Richard Norman, Richard De Abreu Lourenco, Deborah Street, Juliette Malley and Rosalie Viney supported the conceptualisation of the study and the development of the study design. They also contributed to the interpretation of the results and supported the development of the manuscript.

2. **Mulhern B**, Norman R, Street D, Viney R. One method, many methodological choices: A structured review of Discrete Choice Experiments for health state valuation. Pharmacoeconomics. 2019; 37(1):29-43.

   Author Contributions for this paper: Brendan Mulhern conducted the literature search, extracted the data, led the data synthesis and interpretation and developed the first draft of the manuscript. Richard Norman, Rosalie Viney and Deborah Street supported the data extraction process and interpretation of the results and were involved in the development of the manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Economic evaluation is an important tool in health care resource allocation. Interventions are typically evaluated through a cost utility analysis (CUA) using the Quality Adjusted Life Year (QALY), a metric combining length of life and quality of life (QoL) into a single outcome. The quality aspect of the QALY is often provided by a preference-based measure (PBM) that includes a way of measuring health, and a preference-based value set. The most commonly used PBMs focus on health-related quality of life (HRQoL). However, there is a case for broadening what is measured and valued by including other aspects of QoL (such as social care related QoL) alongside HRQoL.

This thesis explores how methods for the measurement and valuation of health and QoL can be extended to inform the development of broader and more widely applicable instruments. This was investigated by first exploring how to incorporate QoL concepts into PBMs, and second by testing the further application of Discrete Choice Experiment (DCE) methods to value QoL. Three empirical studies were conducted

The first study assessed existing measures of health and QoL using Item Response Theory (IRT), and tested two ways in which PBMs could be broadened to incorporate wider QoL concepts. The results demonstrated overlap and divergence in what is measured. This informed where extra dimensions of QoL could broaden the information collected, and how the information collected within existing HRQoL frameworks could be extended.

The second study used DCE to understand respondent preferences for diverse dimensions of QoL. The results provided evidence respondents do trade across different concepts of QoL. This supports the need for broader measures, and also the use of DCE to value broader outcomes.

The third study focuses on DCE methods, and particularly on the construction of designs for DCEs. The results provided detailed information about different design strategies for the valuation of QoL outcomes.

The overall findings raise key issues about what should be captured in PBMs, and also provide novel information about methods that can be used to inform the assessment, development and valuation of future instruments. For example, the results inform how IRT can be used in PBM

development. They also suggest how DCE can be used to value diverse QoL concepts. This can inform the development and valuation of broader measurement systems of QoL outcomes that can increase the scope and enhance the applicability of QALY values used in resource allocation decision making.

LIST OF ABBREVIATIONS

Abbreviation	Description
15D	15 Dimension
A	Anxiety (PROMIS dimension)
ABC	Assessment of Burden of COPD
AC	Accommodation (ASCOT dimension)
AD	Anxiety/Depression (EQ-5D dimension)
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance
AQoL	Assessment of Quality of Life
AQoL-8D	Assessment of Quality of Life – 8 Dimension
ASCOT	Adult Social Care Outcomes Toolkit
BIC	Bayesian Information Criterion
BPI	Behaviour Problems Index
BWS	Best Worst Scaling
CAT	Computer Adaptive Testing
CBA	Cost Benefit Analysis
CF	Crawford Ferguson
CFA	Confirmatory Factor Analysis
CL	Cleanliness (ASCOT dimension)
CO	Control (ASCOT dimension)
COPD	Chronic Obstructive Pulmonary Disease
CREATE	Checklist for Reporting Valuation Studies
CTT	Classical Test Theory
CUA	Cost Utility Analysis
D	Depression (PROMIS dimension)
DCE	Discrete Choice Experiment
DCE_{TTO}	Discrete Choice Experiment including duration
DEMQOL	Dementia Quality of Life
DI	Dignity (ASCOT dimension)
DIF	Differential Item Functioning
EFA	Exploratory Factor Analysis
EORTC	European Organisation for Research and Treatment of Cancer
EQALY	Extended Quality Adjusted Life Year
FA	Fatigue (PROMIS dimension)
FD	Food and Drink (ASCOT dimension)
GDP	Gross Domestic Product
GH	General health (SF-36 dimension)
GMNL	Generalised Multinomial Logit model
HRQoL	Health-Related Quality of Life
HTA	Health Technology Assessment
HUI-2	Health Utility Index – Mark 2
HUI-3	Health Utility Index – Mark 3
ICECAP	ICEpop CAPability measure for Adults
ICER	Incremental Cost Effectiveness Ratio
IIA	Independence of Irrelevant Alternatives
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
IRT	Item Response Theory

LL	Log-Likelihood
LLR	Log-Likelihood Restricted
LLU	Log-Likelihood Unrestricted
LR	Likelihood Ratio
LT-TTO	Lead Time – Time Trade-Off
MAUI	Multi-Attribute Utility Instrument
MH	Mental Health (SF-36 and SF-6D dimension)
MH-RM	Metropolis–Hastings Robbins-Munro algorithm
MIC	Multi Instrument Comparison
MNL	Multinomial Logit
MO	Mobility (EQ-5D dimension)
NICE	National Institute for Health and Care Excellence
OC	Occupation (ASCOT dimension)
ONS-4	Office of National Statistics – 4
OPUS	Older Person's Utility Scale
PA	Pain (SF-36, SF-6D and PROMIS dimension)
PBAC	Pharmaceutical Benefits Advisory Committee
PBM	Preference-Based Measure
PD	Pain/Discomfort (EQ-5D dimension)
PF	Physical Functioning (SF-36, SF-6D and PROMIS dimension)
PICOS	Participants, interventions, comparisons, outcomes, and study
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROMIS	Patient-Reported Outcome Measure Information System
PWI	Personal Wellbeing Index
QALY	Quality Adjusted Life Year
QoL	Quality of Life
RE	Role Emotional (SF-36 dimension)
RF	Role Functioning (SF-6D dimension)
RMSE	Root Mean Squared Error
RMSEA	Root Mean Squared Error of Approximation
RP	Role Physical (SF-36 dimension)
RUT	Random Utility Theory
SA	Safety (ASCOT dimension)
SAS	Statistical Analysis Software
SC	Self-Care (EQ-5D dimension)
SCRQoL	Social Care Related Quality of Life
SD	Standard Deviation
SF-6D	Short Form – 6 Dimension
SF-36	Short Form – 36
SF	Social Functioning (SF-36, SF-6D and PROMIS dimension)
SG	Standard Gamble
SL	Sleep (PROMIS dimension)
SP	Social Participation (ASCOT dimension)
SWLS	Satisfaction With Life Scale
TIF	Total Information Function
TTO	Time Trade-Off
UA	Usual Activities (EQ-5D dimension)
VAS	Visual Analogue Scale

VNM	Von Neumann and Morgenstern
VT	Vitality (SF-36 and SF-6D dimension)
WEMWBS	Warwick and Edinburgh Mental Wellbeing Scale
WHO	World Health Organisation

1. Measuring and valuing health – What are the issues?

Extensive research into the measurement and valuation of health and quality of life (QoL) over the past half century has led to the development of instruments that are used to describe health outcomes, and inform decision making. Common settings for the use of these instruments includes policy development, the measurement of disease outcomes, clinical management, and the economic evaluation of treatments and interventions. An important use of instruments measuring health and QoL is in decision making regarding the allocation of scarce healthcare resources. Economic evaluation has become a crucial tool in health care resource allocation across a range of jurisdictions, with the guiding principle being to choose the interventions that provide the best value for money. This can be measured by assessing the incremental cost per unit of health outcome. Increasingly, this decision making process is informed by using cost utility analysis (CUA) to evaluate new and existing interventions. CUA uses the Quality Adjusted Life Year (QALY) as a key outcome metric. The QALY combines length of life and QoL into a single outcome.

The quality aspect of the QALY is often provided by an instrument that has been developed based on research into the measurement and valuation of health and QoL. These instruments, described as preference-based measures (PBM), are a particular type of outcome measure that incorporate two distinct elements relating to measuring and valuing health. The measurement element of the instrument is a set of questions that are completed by patients or clinical populations to describe their health (known as the health state classification system). This can be generic (for use across all populations) or specific to a particular condition. The valuation element is a way of scoring the health states described that is based on the preferences of a population for the health states described. This scoring system is known as the utility "value set", and is generated using a valuation method that elicits the preferences of a population.

The general concept of QoL is multifaceted and encompasses, and is impacted by, many domains of an individual's life. For example, QoL includes domains of both physical and emotional health, social functioning and relationships, social outcomes, and material aspects such as financial security. Different domains of QoL interact with and influence each other over the lifespan. An example of this is the interaction between health-related quality of life (HRQoL), which includes areas such as mobility and pain, and social outcomes such as safety, control, autonomy and

social participation. HRQoL is impacted by the broader social situation that an individual experiences. As health worsens, and people age, maintaining social outcomes, and the provision of services to maintain good social outcomes, becomes important. Allocation of health care resources across diverse medical conditions, interventions and patient populations is fundamental to achieving better health outcomes for the population. To achieve fair allocation of resources, we need to measure all of the QoL outcomes that matter to the population, and understand population preferences across these different areas of QoL.

Both of the broader research areas relating to the measurement and valuation of health and QoL have different practical, methodological, and normative considerations. These can shape the characteristics of the instruments developed to measure QoL. This includes establishing what should be measured, how this should be generated, and subsequently how what is being measured should be valued. The choices made about these considerations during the development of the instruments can influence their wider use and acceptability by researchers, clinicians and policy makers. Therefore, the characteristics of the methods used to develop and value PBMs, and the influence of different methods on the characteristics of the resulting instruments, requires systematic investigation. This is because there is the possibility that measures will not be sensitive to all of the key aspects of QoL, or may not realistically capture people's preferences. As the instruments are used to inform resource allocation it is important to avoid this and aim to ensure that the QoL changes that are measured in CUA actually relate to things that people care about, and that decisions based on these instruments are rooted in the preferences of the population for which decisions are being made. These issues are returned to in Chapter 2.

There are limitations linked to what is measured, how the measures are developed, and how they are valued. Regarding measurement, generic PBMs have limitations in terms of the concepts (or domains and dimensions) of health and QoL that are measured. This is because they mainly focus on HRQoL concepts at the expense of measuring wider QoL impacts. Therefore, the sensitivity and psychometric validity of these instruments varies across conditions characterised by broader, non-HRQoL focused impacts. This means that the benefits that treatments and interventions may have on broader QoL are underestimated. There are also potential limitations for certain population groups. For example, instruments that focus on general HRQoL factors such as physical functioning, pain, and common mental health conditions may not be sensitive to the benefits of interventions with wider social outcomes in certain

populations such as people in palliative care. Again, effectiveness may be underestimated. Underestimation results in inaccurate QALY values that could bias decision making towards interventions with benefits on HRQoL as represented in commonly used instruments.

Another consideration is that, even in populations where the generic measures are psychometrically acceptable, the information provided by the PBMs is limited. This is because, by their nature, PBMs include a limited set of questions given the requirement that these can describe health states that can be valued. Providing further information linked to dimensions of HRQoL within the existing structure of PBMs is another way to extend their use in a range of settings. Adding further information could provide detailed information relating to each HRQoL dimension. One approach to addressing QoL measurement development issues is through structured application of psychometric methods such as Item Response Theory (IRT) to existing QoL instruments.

The application of methods such as Discrete Choice Experiments (DCE) provide a promising way to address the valuation of health and wider QoL. DCEs are an emerging valuation method that are used for the estimation of utility value sets. However, there are unanswered questions about the most appropriate methodological implementation of DCE, and these knowledge constraints can limit the wider use and acceptance of the methodology. The implementation of the method, and aspects of the protocol used, are influential in shaping the characteristics of value sets. This raises a challenge to ensure that the methods used elicit preferences that reflect the views of the population. One important area of research is to gain a better understanding about how the methodological choices made in the study design process impact on the resulting preferences elicited. This can be done by systematically testing aspects of the DCE study design, and exploring the impact of the design on the value sets produced.

Constraints in knowledge relating to the measurement and valuation of QoL, and limitations of currently available measures, limits their wider applicability. This thesis argues that to there is a case for broadening what is measured and valued to improve the allocation of scarce health care resources. Broadening what is measured and valued relates in this context to including wider aspects of QoL alongside HRQoL in a unified approach. One way in which this could be done would be to combine existing instruments descriptively to form a broader tool, and this has a number of benefits. First, extending the narrow measurement framework could increase measurement sensitivity to the wider QoL impacts of many interventions. This will improve the

sensitivity, and accuracy of values used in resource allocation decision making. Second, if it is possible to combine diverse concepts in the same measurement framework, then research is required to investigate whether instruments combining QoL can be valued on a common utility scale. This has the benefit of using values in decision making that are a result of trading off between diverse concepts of QoL, rather than based on preferences focused on HRQoL. This enables more informed values with a wider applicability across settings, but also values informed by impacts on broader QoL concepts with relevance across diverse health conditions, populations and interventions, to be used. The feasibility of valuing diverse outcomes on the same scale to facilitate the broadening of QoL in a unified approach requires further investigation. It could be argued that diverse outcomes should be considered using different instruments specific to a population or condition. However this would limit comparability across diverse interventions and populations. The values used in decision making would also not be contextualised considering the broader impacts of diverse interventions that may be competing for funding.

1.1. Research question

The arguments described above means that further research is required to investigate how health and QoL are described and measured, with a view to developing a unified approach to measurement. Subsequently, further research is needed to understand the valuation of health and QoL for the estimation of value sets for use in decision making. These research issues are investigated in this thesis.

The overarching question of this thesis asks how methods for the measurement and valuation of health and QoL can be used to inform the development of broader and more widely applicable instruments. This thesis aims to add to our understanding of the application of PBMs to decision making in two ways: first by exploring how broader concepts that are relevant to health and QoL can be incorporated into PBMs, and second by providing new understanding of the application of DCE methods for the valuation of health and QoL.

Regarding the first point, the thesis involves an investigation of the potential to broaden measurement in two ways. First, it investigates whether the dimensions within existing PBMs can be extended to include broader dimensions of QoL. The second investigation assesses whether existing PBM frameworks can be extended to provide further information for each

dimension whilst also providing preference-based information. This will be done using psychometric and IRT methods.

Regarding the second point, the appropriateness of the use of DCEs to value broader measures of health will also be investigated. This work will add to existing knowledge around the extending the measurement and valuation of health and QoL. It will also further develop and test the methods that are used. Combining the evidence from the empirical work will help to understand whether a broader method of measurement and valuation that is useful for decision makers can be developed.

1.2. Aims of thesis

The aims of the research conducted in this thesis are as follows:

1. To conduct a structured review of the use of DCEs for health state valuation;
2. To investigate the relationship between a range of QoL outcome measures assessing different concepts, and the potential for applying these in a broadened measurement framework using dimensionality assessment methods;
3. To investigate the potential for providing further descriptive information within the existing framework of PBM dimensions using IRT methods;
4. To test the use of DCEs to develop a value set for a combined measurement system assessing different QoL concepts;
5. To compare DCE designs that could be used to value wider measurement systems.

1.3. Plan of thesis

The aims described above will be investigated across a structured review and three empirical studies reported across seven chapters. Chapter 2 provides a detailed summary of the theoretical framework within which the thesis is situated. It also introduces key concepts relating to the measurement and valuation of health and QoL. This includes background information, an explanation of the DCE methodology, and a description of the QoL measures used in this area with associated advantages and disadvantages. Chapter 3 describes a structured review of existing literature that uses DCE methods for the purpose of health state valuation (Aim 1). Chapter 4 focuses on the assessment of measures of health and QoL using psychometric and IRT methods, and assesses the two potential ways in which outcome measures could be broadened to incorporate a wider QoL framework (Aims 2 and 3). Chapter 5 focuses on valuation, and describes a DCE study combining two diverse dimensions of QoL to

understand how respondents trade across different concepts (Aim 4). Chapter 6 builds on this DCE work to test a range of methods for the construction of the designed experiment, a key feature of the DCE process (Aim 5). Chapter 7 presents the key outcomes from the empirical work and the implications of the findings, discusses limitations, and proposes a way forward for the measurement of QoL outcomes.

2. Background to the measurement and valuation of health and quality of life

2.1. Chapter summary

This chapter introduces key concepts of relevance to the research conducted. This includes the overall principles that relate to the economic evaluation of health interventions within a CUA framework. This is followed by a discussion of the concept of preferences, the development of utility value sets, and their use in health economics. Different methods used to elicit population preferences and develop value sets, and their advantages and disadvantages, are then introduced. This is followed by a detailed description of the DCE methodology (as the key valuation approach used). The instruments that are currently available to assess HRQoL for use in CUA, and their benefits and limitations are discussed. Measures available for assessing broader QoL are then introduced. Finally, the information provided is linked to the overall research question and aims of the thesis.

2.2. The economic evaluation of health interventions

An individual's health is important, not only to themselves and their family, but also to wider society. There are many determinants of an individual's health. These include their genetic history, lifestyle choices, social situation, and the influence of society and culture on the person. Society as a whole benefit from good health as savings on healthcare expenditure is made, population wellbeing is improved, and productivity also increases. New approaches to improving health and healthcare are therefore important to society, and are constantly being developed. Examples include the development of novel interventions and treatments for different health conditions, new care pathways for population groups with particular requirements, and improved health care facilities. All of these can produce better health

In an ideal world, it would be possible to fund all new innovations, and a major component of healthcare funding is for new treatments. However, the resources available to fund healthcare are limited. The resources available also vary across countries with different health and social care settings and policies. For example, the Organisation for Economic Co-operation and Development (OECD) [1] found that, amongst member countries in 2018, healthcare spending as a percentage of gross domestic product (GDP) ranged from 16.9% (United States) to 4.3% (Turkey) with an average spend of 8.8%. Australia reported spending 9.3% of GDP on healthcare.

The finite resources that are available for health care internationally need to be allocated efficiently. This involves informed choices by health care decision makers, and potentially raises a series of controversial questions. These relate to the type of interventions that should be funded. For example, should interventions that extend a patient's life be preferred to those that improve a patient's QoL? How should interventions that do both in different ways be assessed? A second area of questions can be framed at the population and health condition level. For example, should we fund treatments for common or rarer conditions? We also need to ask questions about which populations should benefit from healthcare resources. For example, should interventions for children be preferred over care pathways for the elderly?

These are important and difficult questions requiring careful decisions. One way in which the decision making is informed is by conducting economic evaluations to assess the 'value for money' of different options based on a measure of the value of the healthcare for both the individual and society. This process is central to healthcare resource allocation decision making in many jurisdictions internationally. In many countries, allocation decision making is conducted by particular agencies who outline guidance about the processes required. These processes differ between jurisdictions, but are generally developed to support the fair and rational allocation of limited resources based on an assessment of cost effectiveness. Examples of international decision making agencies include the Pharmaceutical Benefits Advisory Committee in Australia [2], the National Institute for Health and Care Excellence (NICE) in England and Wales [3], the Canadian Agency for Drugs and Technology in Health [4], and the College voor zorgverzekeringen in the Netherlands [5]. Although each agency specifies different processes, there are some common approaches. This includes the widespread use of CUA as the method of economic evaluation used to compare the benefits of different interventions across different health care types. CUA is discussed in more detail in Section 2.4.

2.3. Key theoretical approaches and concepts important for economic evaluation

The economic evaluation of healthcare interventions can be informed by a number of theoretical approaches based in the wider economics literature. These are the Welfarist and Extra-Welfarist approaches. Both of these theoretical approaches have at their heart the principle of 'utility'.

2.3.1. The utility principle

The key theoretical principle underpinning much of the empirical work presented in this thesis is that of 'utility'. This is rooted in the theory of decision making in uncertain conditions developed by von Neumann and Morgenstern [6]. Utility is also rooted in the philosophical school of 'utilitarianism', where a central tenet is the maximisation of pleasure. According to Seixas [7] "utility relates precisely to the idea of individual satisfaction derived from a given service or good", and to Coast is "the quantity that an individual should maximise or that society should help him to maximise" [8]. In other words, the utility principle specifies that individuals who are rational decision makers will attempt to maximise their utility when making choices, and will therefore choose their preferred option.

The concept of utility as applied in economics is ordinal and arbitrary. Numbers can be assigned to represent utilities as cardinal indicators of preferences. According to Brouwer et al [9], for example, "utility measurement is a systematic method of assigning numbers to entities according to an explicit choice-related rule". The process of understanding both what entities preferences should be numerically elicited for, and how those preferences should be elicited and assigned to represent utility, is a key issue investigated in this thesis. The concepts of utility and utility maximisation are central to both the Welfarist and Extra-Welfarist approaches to economic evaluation.

2.3.2. The Welfarist approach to economic evaluation

The 'Welfarist' approach to economic evaluation is grounded in welfare economics, and supports the maximisation of societal welfare by assuming that some states of living are more preferable to society than others [10]. Individual utility characterises all outcomes, and social welfare is conceptualised as a function of individual utilities [9]. The welfare economics framework relies on four grounding principles for understanding particular states of living as preferable to others [9]:

- Utility: As stated in Section 2.3.1, the utility principle specifies that individuals who are rational decision makers will maximise utility by choosing their preferred option.
- Individual sovereignty: The sovereignty principle states that individuals are able to judge the different factors that are important in contributing to their overall utility and the extent of the contribution can be evaluated.
- Consequentialism: Consequentialism implies that individuals understand that utility is the result of a set of outcomes, rather than the process leading up to achieving the outcomes.

- Welfarism: Welfarism in general states that the utility of a situation is judged by individuals solely in terms of the utility achieved for that particular situation.

There are classical and neo-classical perspectives on these principles. The classical tradition argues that utilities are cardinal, and can be accumulated across individuals. Classical welfarists therefore argue that the optimal situation is achieved when the maximum utility for a particular population is reached. In the neo-classical tradition, utility is perceived as ordinal rather than cardinal, and is therefore more specific to individuals. This limits comparability and means that judgement and comparisons require the use of the 'Pareto principle', which Brouwer and colleagues define as:

> "increase of utility for one individual that involved no utility loss for another was [seen as] an improvement, and an optimum was where no reallocation of resources could be made without reducing at least one person's utility" [9] (pg. 328)

Decision making based on the Pareto principle alone is difficult as there may be many situations where the principle holds. The more flexible Kalder-Hicks criterion builds on what is termed a potential Pareto improvement by specifying that those with an increase in utility from the reallocation of resources can adequately compensate those losing utility as there will still be an overall gain. This is described as the maximisation of total utility.

The Welfarist approach is usually operationalised using cost-benefit analyses (CBA), with health outcomes valued in monetary terms. This approach conceptualises the results as a ratio of benefits and their costs. The net benefit of a one intervention against another can also be considered. It could be argued that this approach is not always favourable for health care, as individual-specific outcomes are central to the assessment of treatment effectiveness. However, the framework is still beneficial for the research conducted in this thesis which investigates the potential to broaden the outcomes that inform resource allocation decision making.

2.3.3. The Extra-Welfarist approach to economic evaluation

The Welfarist approach leads to the Extra-Welfarist perspective which is also a grounding theoretical standpoint relevant to the research conducted in this thesis. This approach aims to maximise the overall health of society in a resource-constrained system. Gyrd-Hansen [11] describes Extra-Welfarism as "not [defining] the output of healthcare in terms of preferences for health vis-a-vis other goods, but according to its contribution to health itself, i.e. they [Extra-

Welfarists] wish to maximise health as against overall welfare." Brouwer, Culyer [9] highlight four key features of Extra-Welfarism. First, outcomes other than utility are possible; second, sources of valuation other than those effected can be used; third, outcomes can be weighted according to non preference-based principles; and fourth it permits interpersonal comparisons of wellbeing across different dimensions. This allows for a different approach to decision making beyond the Pareto principle restrictions than the Welfarist approach, where money is the numeraire, and a strict utilitarian aggregation is imposed. The flexibility of preference-based measurement can be couched within an Extra-Welfarist framework. Incorporating additional flexibility into the Extra-Welfarist framework means that the conceptual approach can be extended to include broader domains of QoL alongside health-related domains in PBMs, and subsequent economic evaluations. Therefore this thesis is grounded in an extended Extra-Welfarist perspective. Building on this further, Extra-Welfarism can be operationalised using a CUA approach.

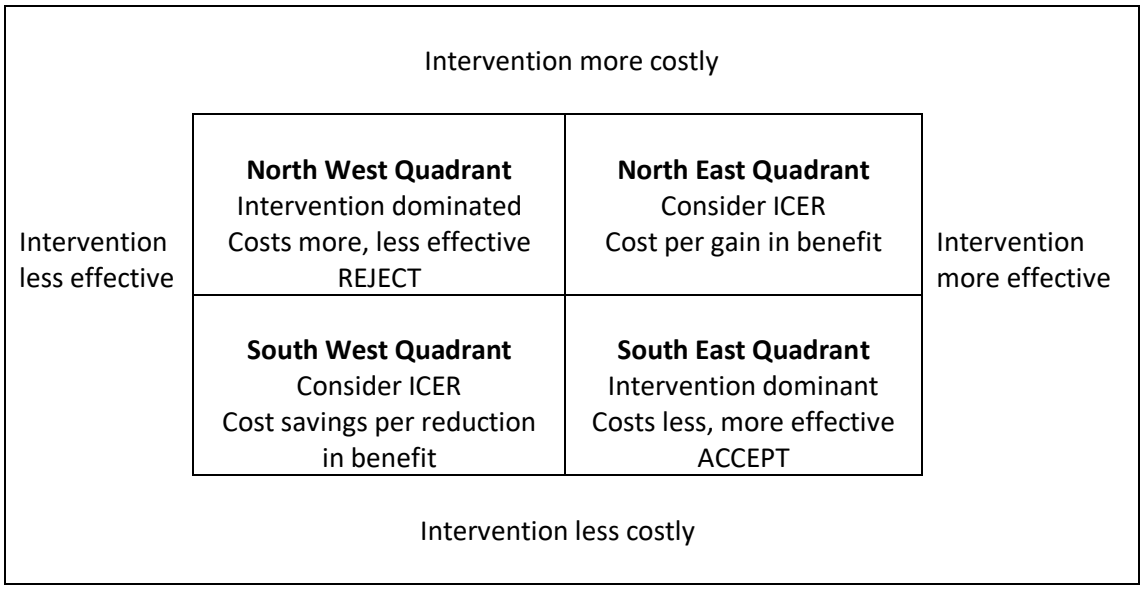
2.4. Cost utility analysis and the Incremental Cost Effectiveness Ratio

CUA is a widely used Extra-Welfarist approach to the economic evaluation of healthcare [12]. CUA is informed by the incremental cost effectiveness ratio (ICER), which allows for direct comparisons of the costs and benefits of new interventions and comparators (which could be, for example, an existing treatment or recommended care). The ICER divides the difference in the cost between two interventions by the difference in their effect. The average incremental cost per unit of an effect indicator is estimated. This effect indicator is health or QoL outcomes, which uses the QALY as the standardised metric (see Section 2.5 for a detailed description). The incremental cost per effect unit of a new intervention or treatment in comparison to an existing one can be compared to a threshold value, where a value lower than the threshold is deemed cost effective. For example, this threshold could be AUD50,000. Therefore, CUA differs to other approaches such as CBA as it compares the benefit of different interventions using the gain in health as the outcome. It aims to maximise the value from health interventions.

The benefits and the associated costs of both the intervention and comparator can be presented using a 'cost effectiveness plane' with four quadrants relating to four outcomes (see **Figure 1**). The Figure shows that the northwest and southeast quadrants have clear outcomes if the intervention is dominated by, or dominates, the comparator respectively. When the intervention is more effective but also more costly (in the northeast quadrant), the benefits

implied by the ICER are assessed against a cost effectiveness threshold. For interventions in the southwest quadrant, the cost saving is assessed in terms of the reduced effectiveness.

Figure 1: Cost effectiveness plane



2.5. The Quality Adjusted Life Year

As described above health outcomes used in CUA to inform the decision making process are often assessed in terms of the incremental cost per QALY gained. Early proponents of the QALY as an outcome for comparing benefits included Fanshel and Bush [13] and Torrance [14]. The QALY is described in detail by Weinstein et al [15], and combines length of life and QoL into a single metric. This is favourable as the outcomes of any technology can be described in terms of life extension (or reduction), and QoL improvement (or decline). The QALY is calculated by multiplying the length of time spent in a health state by an index measure that reflects the QoL value of that state. One QALY is equivalent to one year in full health, and QALY values can be calculated across multiple suboptimal states and variable durations to generate QALY profiles.

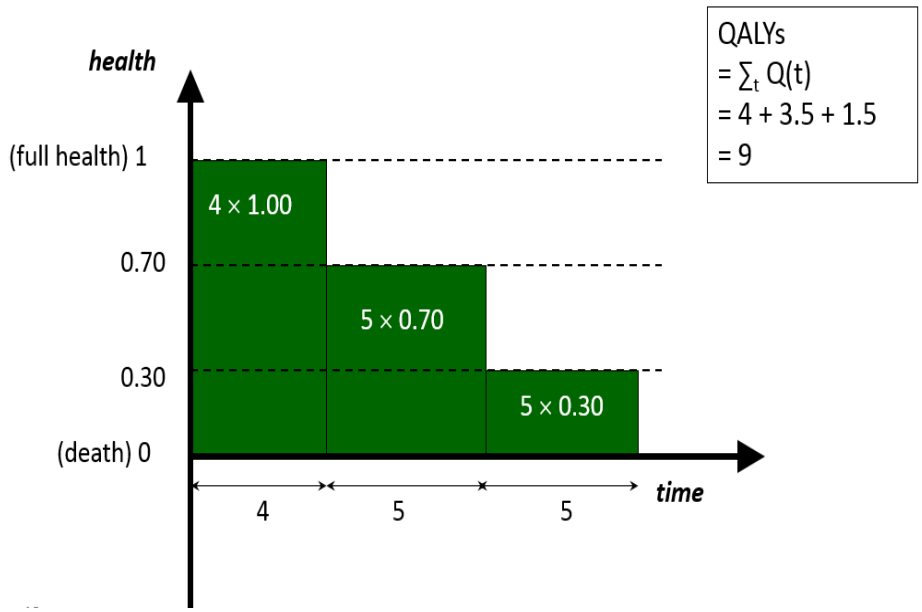
The quality weight of the QALY is known as a health utility value. Health utility reflects the value of a health state weighted in terms of a population's preferences to avoid or achieve that health state. A health utility value is often assigned to a HRQoL state. It is represented by a single index value that can be applied to a health state on a cardinal scale. The value represents preferences for that state. The health states can be taken from many sources, but are commonly

operationalised using the descriptive systems of PBMs. These measures are a key feature of the work conducted in this thesis, and are described in detail in Section 2.11.

The full set of values for a particular predefined set of health states is described as a 'value set' and is anchored on a scale from one (a health state equivalent to full health, or the optimum health state described by a classification system) to zero (a health state equivalent to dead). Negative values are possible, and are equivalent to health states valued by a population as worse than being dead. Given the cardinal properties of the scale, a health state assigned a utility of 0.6 is exactly twice as good as a health state assigned a utility of 0.3. Utility weights for use in the estimation of QALYs are often derived using PBMs that assess HRQoL, but in some cases, broader QoL is measured (see Section 2.11).

Figure 2 displays a potential stylised treatment profile to demonstrate how a QALY profile could be estimated. The horizontal axis represents time spent in a particular health condition. The vertical axis represents the level of HRQoL corresponding to the condition. This could describe the profile of either an individual or a group of patients. In the example given, the level of HRQoL corresponding to the condition decreases over time. The first four years of the profile are lived in a state equivalent to full health (i.e. a health state with a utility value of 1). The next five years are lived in a suboptimal health state with a utility value of 0.7, and the final five years in a more suboptimal health state with a utility value of 0.3. The QALY profile of the treatment is the sum of each time period multiplied by the utility of the health state. In the example this equates to ((1*4) + (0.70*5) + (0.3*5)) = 9 QALYs. An overall profile such as this could then be compared to another health profile to understand changes in QALY profiles between different conditions or treatments. It is the difference between the QALY values that is important, and when comparing two treatments, this can be described as a QALY gain or loss.

Figure 2: Stylised QALY profile



2.6. Health-related quality of life

Health status and HRQoL are widely measured for many reasons. These include understanding the overall clinical profile of patients, assessing the effectiveness of interventions and informing the estimation of health gain and cost effectiveness (most often using QALYs). It is the latter reason that is central to the research reported in this thesis. To measure health it is important to define what is included in health The World Health Organisation (WHO) [16] has stated that health is "A state of complete physical, mental and social wellbeing, and not merely the absence of disease and infirmity". This is a relatively simple but broad definition, and Mayo [17] extends it to describe health in more detailed terms as:

> "A state of complete physical, social and mental wellbeing, and not merely the absence of disease or infirmity. Health is a fundamental human right and is considered a resource for everyday life, and not the object of living. It is a positive concept emphasizing social and personal resources as well as physical capabilities. The prerequisites for health include peace, adequate economic resources, food and shelter, and a stable ecosystem and sustainable resources use." [17]

There are also a number of definitions of HRQoL that vary in the level of detail provided. NICE [3] builds on the WHO definition of health and defines HRQoL as "A combination of a person's physical, mental and social wellbeing; not merely the absence of disease". Leidy et al [18] provide a similar but more detailed definition, saying that HRQoL is ''A person's subjective

perception the impact of health status, including disease and treatment, on physical, psychological and social functioning and wellbeing''. A further definition is provided by Osaba [19], who includes a similar focus on physical, mental and social functioning, but extends it to include other impacts as follows:

> "a multidimensional construct encompassing perceptions of the impacts – positive and negative - of a disease or its treatment on physical, emotional, social, and cognitive functions, as well as somatic discomfort and other symptoms." [19]

Although the level of detail provided in the definitions varies, common themes suggest that HRQoL broadly includes physical functioning (e.g. mobility), social functioning (e.g. leisure activities), psychological health and wellbeing (e.g. depression and happiness), and more widely symptoms of illness (e.g. pain). This means that there is variation in the approaches used to measure HRQoL, and this effects the usefulness of these measures in different settings.

2.7. Measuring and valuing health-related quality of life

As established in the preceding sections, HRQoL is a key component in the estimation of QALYs, and research into both the measurement and valuation of HRQoL is required to facilitate accurate decision making. In section 2.8 the main preference elicitation methods are described. This is followed in Section 2.9 by a detailed description of DCE as the main preference elicitation method used in this thesis. Sections 2.10 and 2.11 describe the most common measures of HRQoL.

2.8. Preference elicitation methods

2.8.1. General principles

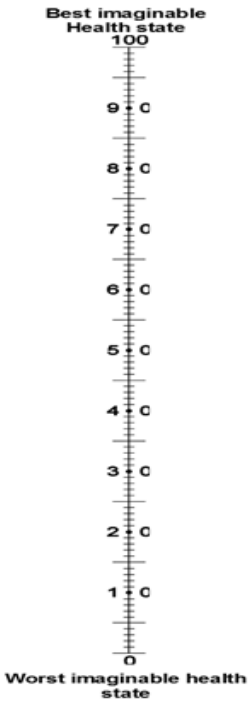
The aim of preference elicitation methods used for the valuation of health and QoL is to estimate the utility of sets of health states. The methods provide a ranking of health states that is quantified in terms of strength of preferences for the states described. In this thesis the focus is on the valuation of states as described by PBMs. This process facilitates the development of value sets anchored on the full health to death utility scale. Valuation methods generally provide respondents with a hypothetical health state, or a number of health states. Respondents are then required to indicate their preference for that state, each set of states, or particular health dimensions. This is done using a variety of approaches with different methodological features that can broadly be defined as rating, iterative, and discrete choice based methods.

2.8.2. Visual Analogue Scale

The Visual Analogue Scale (VAS) is method of health state valuation based on a rating scale approach. It has foundations in psychology as a way to measure responses to sensory stimuli, and is also used as a way to self-report health domains or symptoms (for example level of pain) on a common scale. A VAS is a numbered scale with an anchor at the top and bottom representing best and worst imaginable health (see **Figure 3**). The scales often range between zero and ten, or zero and 100, with further intervals added. Respondents are presented with a set of health states and asked to place them on the numbered scale between the anchors. The intervals between the values allocated therefore reflect differences in preferences for the states. This can result in a score between zero and one for each health state valued. To convert this into a value on the full health to dead scale, respondents also value a hypothetical state described as 'dead'. This is done on the same scale, with values transformed based on the respondent's opinions on where dead lies on the scale. For example, if dead is valued as the lowest then all other states are seen as better than dead, and will have a positive value (and are rescaled depending on the value given to dead if it is not zero). If there are states valued below dead, then negative values can occur.

VAS has been used to value hypothetical health states internationally [20-22]. The advantage of VAS is that it is simple to administer, and it is easy for respondents of a range of ages and sociodemographic groups to complete. However VAS has been criticised for not involving choices, and therefore not capturing a respondent's strength of preference for certain sets of states [13]. This means that it is unclear whether respondents are interpreting the VAS as a cardinal scale as is required for QALY weights. VAS is also prone to context effects including 'response spreading' of the values assigned to states across the full scale presented [23], and also 'digit preference' for the round numbers displayed on the scale. This means that respondents may be more likely to place a state at a value of, for example, 70 rather than 72. These are forms of 'framing effect' based on features of the task [24]. In support of VAS, Parkin and Devlin [25] argue that it does have a theoretical basis in line with CUA, and that an element of choice is involved.

Figure 3: Example VAS scale



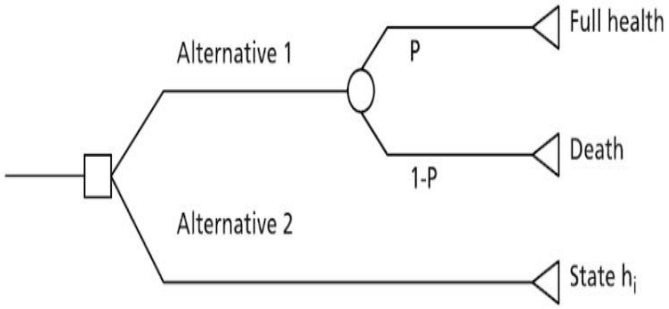
2.8.3. Standard Gamble

The Standard Gamble (SG) is based on the von Neumann and Morgenstern (VNM) theory of Expected Utility, which states that under conditions of uncertainty, a decision maker will choose the option with the highest utility to them [6]. SG is an iterative approach that involves a choice between two treatments with different outcomes, one of which is certain and one of which is uncertain (see **Figure 4**). Alternative one is uncertain as it involves risk, and offers a return to full health for t years (prob. p) OR immediate death (prob. 1 – p). In contrast, alternative two is certain and involves living in the health state that is being valued for t years. The probability of returning to full health is iteratively changed until the respondent is indifferent between alternative one and two. The utility of the health state is equal to the probability of full health at the point of indifference. For example, if indifference occurs between the fixed health state, and a profile consisting of a 0.7 probability of returning to full health and a 0.3 chance of death, the utility value for that health state is 0.7.

SG has an established theoretical basis, although individuals may not conform to the assumptions of VNM theory. There are a number of other concerns with the approach, some of which apply to other valuation methods. First, respondents can find probabilities difficult to understand and evaluate [26], and may overvalue small probabilities and undervalue high probabilities [27]. Second, the tasks are complex, and some groups of respondents will not have

previously considered risk in terms of health gain and loss. However, as a counterpoint, other groups of respondents will have made risk based decisions for healthcare or treatment. This may lead to different levels of understanding, and face validity of responses. Finally, people are generally risk averse and desire to avoid death, so do not accept a high probability of dying. This can lead to higher utilities for severe states in comparison to those produced from other valuation methods [28]. This could impact on decision making depending on the values that are used.

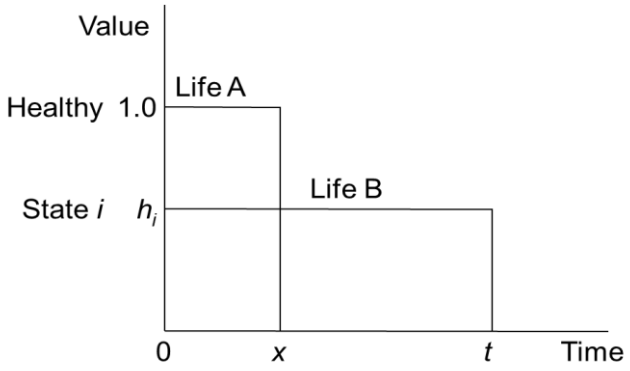
Figure 4: The Standard Gamble valuation process



2.8.4. Time Trade-Off

The Time Trade-Off (TTO) elicits preferences for health states by asking respondents to trade time rather than risk. TTO is an iterative valuation process that produces health state values by asking respondents to trade years of life in a state defined as 'full health' to avoid a fixed time in a state describing a suboptimal health state.

Figure 5: The Time Trade-Off process



For a diagrammatic example of a TTO task, see **Figure 5**. Respondents choose between a set amount of years (*t)* in a health state h_i (Life B) and *x* years (which is between 0 and *t*) in full health (Life A). The most commonly used value of *t* is 10 years. The amount of time in full health in life A is varied following an iterative process until indifference between the lives is reached. The

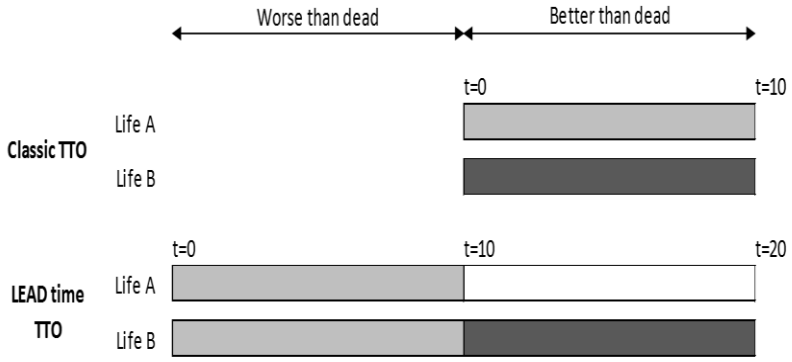
value for that health state is then calculated as x/t, which can be rescaled onto the one to zero scale dependent on the value of *t*. Health states can also be valued worse than dead. For example, take the situation where *x* = 0 years in full health in life A (equivalent to immediate death) is preferable to 10 years in h_i). In these cases, a number of approaches can be used to value the states. In earlier research valuing states worse than dead [29], respondents choose between w years in h_i followed by x years in full health (where $w+x=t$, with *t* often set at 10), and immediate death. The length of time in h_i is iteratively varied until the indifference point is reached. However, using this task for the valuation of states worse than dead is a methodological concern. It produces results on a different scale to the task for states better than dead, which requires rescaling. The frame of reference of the profiles described also differs as it includes a fixed state of immediate death, and a profile which includes a transition between different health states. This means that it requires a cognitive completion process that is not fully comparable to that required for the better than dead valuation task.

An alternative method that can be used to value all states using the same task, and hence avoid the issues with using two different approaches, has been proposed by Robinson and Spencer and Devlin et al [30, 31]. This is known as the 'Lead Time' TTO approach and involves adding an extra 'lead time' in full health to each option. For example, imagine that a lead time of 10 years added to each state, and t is fixed at 10. Respondents then choose between 10 years in full health followed by 10 years in the health state being valued (Life B), or between x years (that can take a value between 0 and 20) in full health (Life A). This means all states can be valued on the same scale, which in this case has a range between -1 (the worst possible value) to 1 (the best possible value that is equivalent to full health). **Figure 6** presents a diagrammatic example comparing the 'classic' approach to TTO with the 'Lead Time' approach.

The TTO has conceptual and methodological issues that can limit interpretation of the values produced. First, iteratively trading life years in full health to avoid a particular health state described using different health concepts and levels is a complex process. Second, in everyday life, respondents do not make decisions involving trading years of life, and therefore there are concerns regarding task realism. Both of these issues can lead to response error. Third, respondent uncertainty regarding the task completion process, and potential burden, may lead to values that lack validity. For example, respondents may provide values that are approximate rather than precise and considered. Fourth, many studies have used a fixed time period of ten years assuming constant proportional trading with respect to time. This means assuming that

the same proportion of time would be traded off irrespective of the overall length of life offered. There is evidence to suggest that this assumption does not hold [32], and there are suggestions that preferences for time are non-linear [33]. Fourth, loss aversion [34], where choices are made in comparison to some reference point, may lead to response bias. Finally, the iterative process used during the study design process can influence the descriptive characteristics of the results, with clusters of values at certain points based on the sequence used [35].

Figure 6: Representation of the Lead Time TTO process



2.8.5. Discrete Choice Experiments

In response to some of the concerns about the complexity of iterative valuation methods, there has been interest in using DCE methods for health state valuation. DCE is an ordinal choice based method built on Random Utility Theory (RUT). RUT assumes that when faced with a multiple profiles or scenarios describing a good or a service, individuals will choose the option that they believe provides them with the greatest utility. In line with Lancastrian choice theory [36], DCE decomposes overall descriptions of a good or service into particular attributes, each of which consists of a number of levels. Following RUT, it is expected that the combination that gives the highest utility will be preferred by respondents. Choices over pairs of profiles are modelled to quantify the impact of attribute level changes on choice. The estimates of the magnitude of these impacts are analogous to the strength of preferences for changes in attributes and attribute levels. In health care, DCE choices could be between particular treatment profiles or care situations. For health state valuation, choices are made between health states that are described in terms of their attributes (widely termed dimensions in the literature), and associated severity levels. It is this approach to valuation that is applied and tested in this thesis.

In the most commonly used approach to DCE for health state valuation, respondents are presented with choice sets (or choice tasks) consisting of sets of health scenarios, and are asked

to choose between them. **Figure 7** and **Figure 8** each present an example DCE choice set format. The health states are often taken from generic measures of health and QoL that are described in Section 2.11. DCE choice sets present respondents with a series of different descriptions and asks them to make a choice between the options. Each respondent completes a set of choice sets selected from an underlying designed experiment. The choice data elicited are aggregated over many choice sets including different combinations of dimension levels, and respondents, and modelled using regression methods (based on McFadden's utility theory [37]) to infer which levels of each health state dimension are preferred by the overall sample. These are expressed as utility decrements associated with each level of each dimension in comparison to the baseline level.

Figure 7: DCE choice set example 1

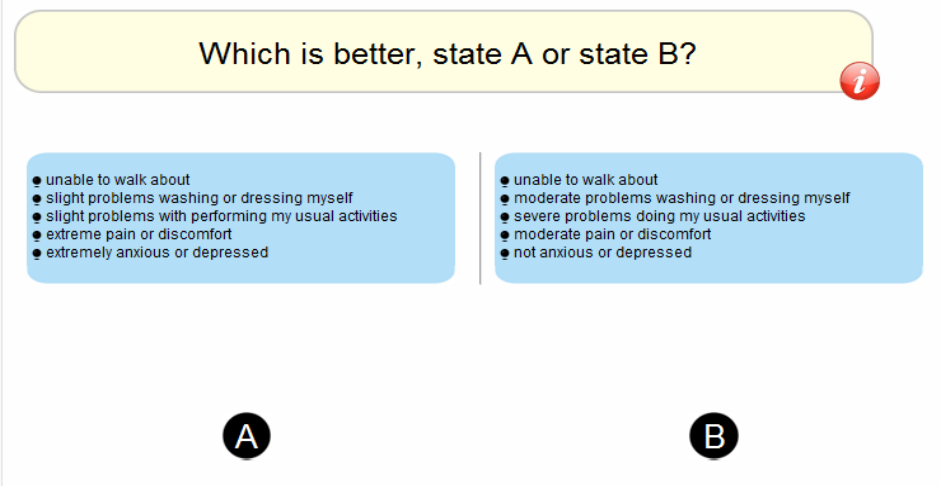


Figure 8: DCE choice set example 2

	HEALTH SCENARIO A	HEALTH SCENARIO B
	Slight problems in walking about	Moderate problems in walking about
	Slight problems washing or dressing yourself	No problems washing or dressing yourself
	Slight problems doing your usual activities	Slight problems doing your usual activities
	Severe pain or discomfort	Moderate pain or discomfort
	Moderately anxious or depressed	Moderately anxious or depressed
	Live for 5 years and then die	Live for 1 year and then die
Which scenario do you think is better?	☐	☐

The key elements of the design, implementation and analysis of DCE studies for health state valuation are described in Section 2.9. There are also a number of published guidance papers for the conduct of DCE studies including widely cited and influential work by Johnson et al [38] and Bridges et al [39]. The use of DCE for health state valuation has expanded in recent years. This has led to the development of different methodological approaches and study designs within the wider DCE framework. The methodological approaches used can influence the characteristics of the value sets produced. However, it is unclear which DCE based methods, and the methodological choices made, produce value sets that most accurately reflect the preferences of the population sampled. This thesis includes a structured review of studies using DCE methods for the purpose of health state valuation (Chapter 3). Section 2.9 builds on the brief discussion of DCE outlined in this section to discuss the background and methods of the approach in more detail. This is done as DCE is the methodological approach applied and tested in two of the empirical studies reported.

2.9. Description of the DCE methodology

DCE methods have become widely used in general health economics research (for reviews see DeBekker-Grob et al [40] and Clark et al [41]). The use of DCE for health state valuation is a subset of this work, and has a number of methodological issues. Although the overall DCE framework can be similar across studies, there are choices that need to be made, for example in study design and administration, and the analysis undertaken. Given the diversity in approaches there are methodological issues and choices linked to each stage of the DCE process. These include the development of the descriptive system, the choice set format, the construction of the designed experiment, the sample size calculation, the implementation of the study, and the methods used for data analysis and modelling. **Figure 9** and **Figure 10** present stylised diagrams that are relevant to the DCE development process sources from the guidance papers published by Bridges et al [39] and Johnson et al [38] respectively. Both flowcharts start with the identification of the research question or objectives before moving into the development of the attributes and levels. This is followed by the experimental design phase (including task construction), before moving onto data collection, analysis and reporting. Each of the key steps for the development, administration and analysis of a DCE will be discussed in the remainder of this section.

Figure 9: Conjoint analysis development process

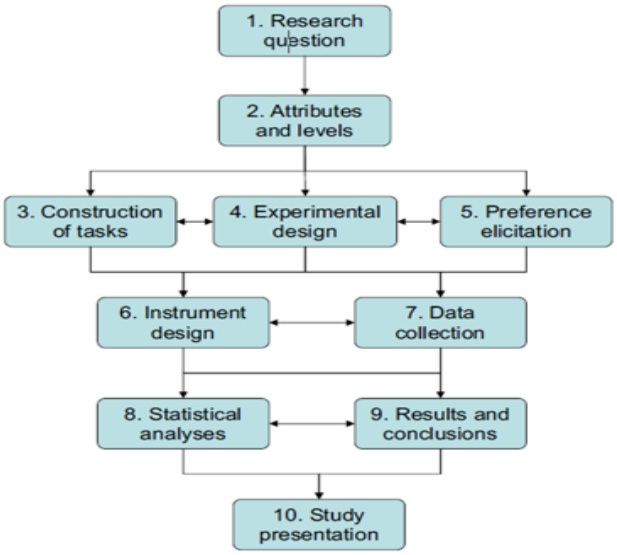
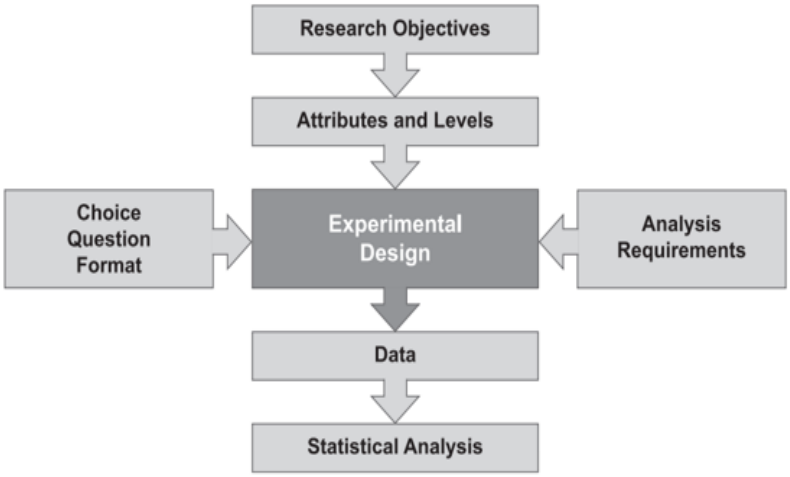


Figure 10: DCE study design process



2.9.1. Developing descriptive systems for valuation

The process of constructing a set of attributes and levels for valuation using DCE depends on the aims of the study, and what is being valued. Outside of the health state valuation research area, attributes are often identified using literature review, detailed qualitative work, and a subsequent refinement process (see Coast et al[42]). In using DCE methods for the generation of utility value sets, descriptive systems are based on PBMs that have been developed using qualitative or psychometric methods, or both. The requirements of the valuation study are often considered during the PBM development process, and this means that the dimension and severity level descriptors are available for valuation. However, some level of adaptation or refinement to ensure that the descriptions are amenable to valuation using DCE can be

undertaken.

The perspective used to frame the profiles is also a consideration, and there is a choice to be made about whether the dimensions are presented without an associated perspective, or in the first, second or third person. In this thesis, the third person perspective is not considered further as the main focus of interest is on the valuation of QoL for an individual rather than using a proxy framework. The first and second person perspectives are the most widely used in health state valuation research. The second person perspective is used in the DCE's implemented in this thesis, as it allows the task to be framed as impacting the respondent by asking them to imagine that "you are living" in particular health states. The dimension descriptions are adapted directly from the PBMs that are valued in the empirical work.

2.9.2. Constructing the choice sets – General format

Another feature of the study design is the development of the choice set format. This includes both the overall layout and the presentation of the attributes. Considering the format is important as it can help support the completion of the choice sets by respondents. However, it is also worth noting that particular formats may influence the decision making strategies used to assess and complete the tasks. A formatting choice that is important in the look and feel of the choice sets is the use of highlighting to distinguish between dimensions. In past studies this has been done in a number of ways, for example by shading alternate dimensions within the choice set, or shading the dimensions where the severity levels differ [43, 44]. Respondent feedback suggests that shading supports the task completion process.

Another key issue is the use of severity level 'overlap'. This is defined as specifying in the design that a certain number of dimensions from the overall system have the same severity level. For example, Jonker et al [44] specified that any two of the five EQ-5D-5L dimensions were to be fixed at the same severity level across all choice sets in the design. The benefit of overlap is that tasks are easier for respondents to complete as there is less information changing within choice sets, and therefore less information to process. The disadvantage is that imposing overlap lowers the statistical efficiency of the designed experiment. Therefore, choosing to use overlap requires a trade-off between the ease of respondent completion, the statistical efficiency of the constructed design, and the overall number of choice sets that are required to estimate the model parameters with adequate precision.

The order in which the dimensions are presented within choice sets is another methodological consideration. This is because presenting the dimensions in the same order could lead to bias towards certain dimensions based on completion strategies that focus on the first or last dimension presented as a short cut to completing the tasks. There are varying levels of health state dimension ordering that could be imposed including using a fixed order, moving subsets of dimensions, or allowing all possible dimension orders within a descriptive system to be presented with variation either between respondents (i.e. one respondent sees one order) or within respondents (i.e. each respondent could see a different possible order for each task). Previous research testing the impact of dimension order effects on responses has proved inconclusive [45, 46]. Therefore, the level of dimension order randomisation to implement is a methodological choice by the researcher.

A further general format issue linked to choice set construction is the question asked of respondents. In health state valuation DCE work, the most commonly asked questions are which health state is 'best' [47] or which they 'prefer' [48]. Depending on the number of profiles presented within choice sets, respondents have also been asked to choose the options that they feel are best and worst.

There is also a decision to make regarding the number of options to present to respondents. This links to the overall aim of the study, and the possible requirement for anchoring onto the full health to death utility scale (for more information see Section 2.9.3). For example, pairs of health states are commonly used [47, 49], and triples have been presented as three options [50], or as two sets of pairs[51]. Quadruple profiles have also been presented to respondents [52].

In the DCE studies conceptualised, designed and implemented in this thesis, choice sets are constructed as pairs of QoL states presented in either a fixed order, or randomised at the subset level. The dimensions involve highlighting of the key severity levels, and overlap across dimensions is imposed. The question asked of respondents is to indicate which QoL state they prefer.

2.9.3. Constructing the choice sets – Anchoring

An important issue for the wider application of DCE methods for health state valuation is that the utility values produced are estimated on a latent scale (so the estimates could take any value), and are therefore not anchored on the full health-dead utility scale. In response to this

issue a number of solutions have been proposed. One such solution has suggested that latent DCE values are anchored using external preference data from another valuation method such as health state ranking, or TTO [53, 54]. This approach has the advantage of combining preference data elicited using different, but complementary, approaches, potentially generating more informed values. Disadvantages of this approach relate to potential learning effects based on the overall task order, data quality issues, and the need for modelling strategies that can combine the data.

An alternative solution that has gained traction is the inclusion of duration as an attribute in the DCE choice set (DCE_{TTO}, [48, 50, 55-58]). This approach allows for values to be directly modelled onto the utility scale. This method was simultaneously developed by multiple research teams, and generates health profiles consisting of QoL state descriptions for an associated duration [48, 50, 57]. The choice data is modelled using regression methods, with the estimates 'anchored' by the duration coefficient where each health state dimension level is divided by the duration parameter estimate.

Within the DCE_{TTO} framework there are further choice set construction issues that can influence the anchoring of the values. For example, the design of the duration attribute is important, and there are considerations required regarding the number of duration levels included, the value of those levels and the range that they cover. Increasing the number of levels can improve the precision of the estimate of the duration parameter, but including many levels can increase the complexity of the design. Choosing the duration levels so that the values within a choice set vary from being closely matched to presenting values with a larger difference facilitates trading on both the health state dimension levels and the duration attribute.

Another important choice set framing decision for the implementation of DCE_{TTO} is the use of pairs of states, or the inclusion of a third profile in the overall choice scenario. The approach developed by Bansback et al [48] and later implemented internationally [55, 59] presented pairs of health states including duration. Simultaneously, Viney et al [50] and Norman et al [57, 58] developed an approach that included pairs of states with an associated duration, and a third option describing immediate death. Under this framework, respondents choose the best and worst option of the three. More recently, Jonker et al [60] have developed a format that includes 'full health' for a shorter duration than the non-optimal states as the third option.

In this thesis, the DCE studies reported are not designed to incorporate duration. This is because the studies are methodological investigations of particular features of the DCE valuation process, and are not conducted to develop value sets for use in the estimation of QALYs.

2.9.4. Constructing the experimental design

Implementing a DCE requires the construction of a designed experiment (i.e. the selection of choice sets to include in the study). There are two broad classes of DCE design construction that can be described as theoretical-based and algorithm-based [61]. These are described below.

A type of theoretical-based method is the generator developed approach, which was created by Street and Burgess [62]. To construct a design using this approach, an orthogonal array [63] is taken as the starting point, and each row of level combinations within that array is used as the first health state in a choice set. The second option is constructed by making a systematic set of level changes given by the chosen generator so that each level of each dimension appears as evenly as possible across the options in each choice set (assuming that main effects are to be estimated).

In algorithm-based approaches, a starting design of choice sets is selected, and this is improved on to generate a more efficient design. This can be done by changing one profile at a time (a modified Fedorov algorithm), or by changing dimension levels within profiles (a coordinate exchange algorithm) [64, 65]. A further variation with the algorithmic design construction is the software used to generate designs in practice. In the field of health state valuation, several programs have been used, including routines available in statistical software such as Stata [66] and SAS [65], or using and specifically designed DCE software such as Ngene [67]. The impact of the different software implementations on the designs produced is unclear.

There is also the potential to include non-informative or informative prior values for each dimension or dimension level in the design process. Non-informative priors take the value of zero, and therefore do not provide the design process with any information about what the magnitude of the attribute level should be. Informative priors take non-zero values, and provide the design with information about the magnitude of each dimension level. As with other design choices there is a trade-off between the ease of completion for respondents and the statistical efficiency of the design. With regards to health state valuation, there have been limited comparisons of the different types of designs, both in terms of their statistical efficiency and

actual completion by respondents. It has been suggested that designs constructed using informative priors are more difficult for respondents to complete [59]. However, further work is required to understand how different priors influence the utility values obtained.

2.9.5. Implementing the DCE – Task allocation

Task allocation within a DCE study is closely linked to the designed experiment. When generating a set of choice sets, a decision needs to be made about how to allocate the tasks to respondents. One option is to group the choice sets into 'blocks', and randomise respondents to complete one of the blocks. A second option is to randomly select choice sets from the overall design. Blocking can be performed as part of the choice set construction process, and aims to ensure that a selection of dimension severity levels and combinations of these (and overall severities) are included in each Block. The choice sets within each Block are presented to respondents in a random order. This means that each respondent will make choices based on a wide range of severity level trade-offs. However, a potential bias could be introduced if a particular Block of choice sets is not balanced. Random allocation does not have the same level of control over what choice sets a respondent completes. So a respondent could complete a group of choice sets with a particular bias towards certain severity levels. However, at the overall level, the randomisation process should ensure a relatively equal number of observations across all choice sets.

Alongside the presentation of the choice sets, extra tasks are often included to understand respondent comprehension and level of attention. For example, a choice set can be repeated at multiple points of the data collection to understand response consistency. If the answer differs then it may be an indicator of lack of attention. It is also possible to include dominated choice sets, where all levels of one profile are logically better (that is, at a less severe level) than the other. The choice made is an indicator of respondent processing of the tasks. Johnson et al [68] provide a summary of the use of internal validity consistency checking questions in DCE studies, and found that internal validity is rarely reported.

Multiple study design approaches have been used in the implementation of DCE studies for the valuation of health states, and the approach used depends on the aims of the study. For the development of value sets, single arm studies will often be used. For methodological work, parallel arms studies collecting data across multiple formats or designs, and crossover studies (with respondents completing more than one arm or design) have also been utilised.

In this thesis, single and parallel arm DCE studies are favoured. For choice set allocation, both blocking of tasks within the design, and random allocation without replacement are used. A repeat consistency checking question is also implemented in one of the empirical studies to assess respondent completion patterns.

2.9.6. Implementing the DCE – Sample size, choice set and observation numbers

The sample size required for studies is also an important consideration. Sample size is linked to the number of observations required per choice set which in turn is linked to a number of other considerations. These considerations include the complexity of the models to be estimated, the overall number of choice sets in the design, and the number of choice sets competed by each respondent.

The overall number of choice sets to include in a DCE design is not determined by any clear criteria other than the requirement for the data collected to be sufficient to estimate the parameters in the model. One criterion used in the health state valuation DCE work suggests that the number of choice sets in the design should at least exceed the number of parameters there are to be estimated in the model. The overall number of choice sets is also linked to the sample size available, and the number of observations that can be obtained per person. There is a trade-off between collecting more observations for each choice set in a smaller design, or fewer observations on choice sets within a larger design. Having access to a large sample may mean the number of choice sets each respondent is required to complete could be lower than if the sample is smaller, and this could maintain data quality and attention levels. The number of choice sets an individual can complete is associated with the complexity of the choice sets presented. Further information about the sample sizes and number of choice sets used in existing health state valuation studies is reported on as part of the structured review reported in Chapter 3.

In the empirical DCE work reported in this thesis, sample and design size were interpreted in terms of the numbers of observations overall, per choice set, and per person. The number of parameters to estimate in the models, and the wider aims of the studies, informed the design approaches.

2.9.7. Implementing the survey – Mode of administration

DCE studies have been administered using a number of different modalities. These include pencil and paper, postal distribution, interviewer administered both face-to-face and via phone, and online. Online methods facilitate the collection of data from large samples of the general population who may be more difficult to reach using interview focused methods. Online data collection has the benefit of reaching large and representative samples relatively quickly and cheaply, and with good response and completion rates. However, a disadvantage of online methods is a lack of control over the environment in which the survey is completed. To counteract this, indicators such as the time taken, and feedback questions about the task completion process, are used to attempt to understand respondent engagement.

Postal distribution of surveys has many of the same concerns as online data collection, but also results in a lower response rate and can be more time consuming for respondents and researchers. It is also less flexible for the implementation of key methodological considerations such as randomisation. Face-to-face interviewer led valuation data collection is also time consuming and can be expensive. However, the interviewer can exert increased control over the interview process and environment, and intervene if the respondent is having difficulties, or there is a lack of engagement.

A noteworthy point about respondents recruited for online, postal and face-to-face data collection is linked to the representativeness of the sample. Samples recruited via any process can be matched with the general (or a patient) population in terms of observable characteristics such as age, gender or region. Surveys can also include a range of demographic and other exploratory questions linked to the topic (for example self-reported health and experience of health in themselves and their families). However, there may be unobservable characteristics linked to those individuals recruited via different modes that are difficult to measure.

In the DCE studies reported in this thesis, online data collection was used. This is because it allows for large samples of representative respondents to be recruited. Online methods are also widely used to collect choice data in Australia as for geographical reasons, data collection from representative samples using interview based methods is difficult.

2.9.8. Implementing the DCE - Survey format

Online health state valuation DCE surveys generally include some consistent elements. After entering the survey, respondents are provided with information about the background and aims

of the survey. There is also a consent process, which may be implied consent by accessing the survey and clicking past the information to commence the questions, or a page formally gaining consent by agreeing to take part in the survey. Demographic quota questions are included to ensure that a particular respondent belongs to a quota group (often based on age and gender) that is still required. If a particular group is complete, then respondents in that group are not usually able to complete the survey.

More extensive demographic questions are also often included to compare the sample to a matched population. Respondents may also complete the descriptive system they are going to value for their own health and QoL. This provides information about their own perceived QoL whilst also familiarising them with the dimensions included in the DCE task. Surveys also often include an instruction page, and an example task to helps respondents to understand the requirements of the study. Further questions specific to the topic of the DCE, or feedback questions about the tasks, may also be included after the choice tasks have been presented. Free-text questions allow for the provision of further unrestricted information, and these are useful to identify any issues with the content and functioning of the survey.

2.9.9. Data analysis and modelling

There are various methods for the modelling of DCE data. In order of increasing complexity, these are the conditional logit, scale assessment, latent class, mixed logit, and generalised multinomial logit models. Each model has different underlying assumptions and features that allow for a detailed understanding of the preference patterns of a particular sample. These models are tested in the two DCE studies reported in this thesis.

2.9.10. Conditional logit model

The conditional logit model is a widely used implementation of the multinomial logit model for the analysis of DCE data. The model conceptualises RUT [69], and assumes homogeneity of preferences across the population sampled. Coefficients within dimensions (i.e. estimates of the model parameters) are interpreted as decrements or increments on a latent scale in comparison to the baseline level. A larger change is indicative of a higher preference to avoid or accept that level in comparison to the baseline. Conditional logit models also assume independence from irrelevant alternatives (IIA) which states that a choice between two options is independent of any other choices that could be made [70]. Equation 1 describes the utility function for the conditional logit for individual ***i***, with ***j*** alternatives in scenario ***s***:

$$\boldsymbol{U_{ijs} = X'_{ijs}\beta_i + \varepsilon_{ijs}}$$ (1)

where $\boldsymbol{\beta_i}$ and $\boldsymbol{X'_{ijs}}$ are the vectors of the coefficients and explanatory variables respectively, and $\boldsymbol{\varepsilon_{ijs}}$ is the error term linked to error in the choices made by respondents. The conditional logit is often the initial model tested to understand general patterns of preferences for the overall sample. This can then be developed into further tests of group differences and preference heterogeneity.

2.9.11. Scale assessment modelling

The scale assessment modelling approach was proposed by Swait and Louviere [71], and examines subgroup differences in preference patterns. The approach tests the null hypothesis that the underlying scale (of pattern) of respondent preferences is not different across subsamples (for example between demographic groups). This is implemented by comparing conditional logit models using a likelihood ratio (LR) test as in Equation 2:

$$\mathbf{LR = -2(LL_R - LL_U)}$$ (2)

where $\mathbf{LL_R}$ is the Log-Likelihood of a conditional logit model that is estimated on the pooled sample and allows for scale differences based on a subsample within the data, but assumes that the parameter estimates are the same across the pooled data. $\mathbf{LL_U}$ is the sum of the log likelihoods of conditional logit models estimated on each subsample separately. Combining these models forms an unrestricted model that allows for variation in overall respondent preferences. The LR statistic is then used for comparison. The null hypothesis (of homogeneity) is rejected if the LR statistic is greater than the critical value from a Chi Square distribution. The critical value used is linked to the degrees of freedom in the model. The degrees of freedom equate to the difference in the number of parameters between the unrestricted and restricted models.

2.9.12. Latent Class model to assess heterogeneity

Latent class models identify groups (or 'classes') of respondents in the sample that have similar underlying patterns of preferences. For latent class modelling, the conditional logit model is adapted to allow for heterogeneity at the respondent level ($\boldsymbol{i}$):

$$\boldsymbol{u_{ij} = \beta_i + \lambda'_i x_j + \varepsilon_{ij}}$$ (3)

This produces parameter estimates for different classes of preference patterns within the sample. The number of classes to estimate is pre-specified by the analyst. The usual procedure is to test a series of models with different class numbers, and determine the optimum number of classes using one of a number of model performance indicators. These include the Akaike

information criterion (AIC) and Bayesian information criterion (BIC) [72]. See Section 5.4.18 for a detailed description of these indicators. It is also possible to test the likelihood of certain demographic groups belonging to each class. This is done by estimating the probability different demographic groups belonging to different classes in comparison to a baseline class.

2.9.13. Mixed logit model to assess heterogeneity

In comparison to latent class, which estimates heterogeneity by drawing out groups with similar preferences, mixed logit [73] is a random parameter model that allows an assessment of unobserved preference heterogeneity at the individual parameter level. The mixed logit model extends the conditional logit approach by allowing the parameters in the model to have either heterogeneous (random) or homogenous (non-random) preference patterns. Choosing the attributes to specify as heterogeneous or homogeneous is based on assumptions and hypotheses about the preference patterns of the sample. Equation 4 displays the mixed logit model. In this model, the utility for individual *i* associated with choice *j* in scenario *s* is:

$$\boldsymbol{u_{ijs}} = \boldsymbol{\beta x'_{ijs}} + \left(\boldsymbol{\eta_i x'_{ijs}} + \boldsymbol{\varepsilon_{ijs}}\right) \quad (4)$$

where $\boldsymbol{\beta}$ is a vector of coefficients and $\boldsymbol{X'_{ijs}}$ is a vector of the explanatory variables and $\boldsymbol{\eta_i}$ is a variability term that varies between respondents. The standard deviation for each parameter is estimated, along with the associated significance level. Significant standard deviations for a parameter indicate that preferences for the dimension or attribute level are heterogeneous. Mixed logit models can specify a range of assumed underlying distributions for the parameters. The most commonly assumed are normal and log-normal distributions, with either correlated or independent dimensions. As the likelihood function does not have a closed form, simulated maximum likelihood must be used to estimate the parameters, and the accuracy of the estimates depends on the number and placement of the values at which the likelihood is approximated.

2.9.14. Generalised Multinomial Logit Model to assess heterogeneity

The Generalised Multinomial Logit Model (GMNL) increases the complexity of the mixed logit model by modelling variability at the within and between-person level [69]. This means the model accounts for both scale heterogeneity within the sample, and random parameters that can be specified as heterogeneous. The GMNL model is specified for individual ***i*** associated with choice ***j*** in scenario ***s*** as in Equation 5:

$$\boldsymbol{U_{ijs}} = (\boldsymbol{\sigma i \beta} + \boldsymbol{\eta i})\boldsymbol{X' ijs} + \boldsymbol{\varepsilon ijs} \quad (5)$$

where $\boldsymbol{X'_{ijs}}$ is a vector of explanatory variables and $\boldsymbol{\sigma_i}$ is the scale parameter. The scale parameter

is assumed to be log-normally distributed with its mean equal to 1 [74].

2.10. Instruments used to measure health and quality of life – Profile measures

The DCE methods described above can be applied to generate utility value sets based on the preferences of the population for health states described by PBMs. The other key area for investigation in this thesis is how to measure health and QoL. There are two broad groups of instruments used for this purpose. PBMs are one such group. The other group can be described as profile measures of QoL.

Profile measures include ordinal scales where the scores produced typically have a clinical interpretation. Both generic and condition specific profile measures are available. They are often developed using qualitative input and psychometric methods, and consist of sets of items that measure broad domains of health and QoL. Both item level and domain level scores can be calculated. An example of a profile measure is the SF-36 [75], which is the most widely used generic profile measure of HRQoL internationally, and has been applied in many health conditions [76]. It produces eight dimension scores from 36 items (physical functioning (PF), role physical (RP), role emotional (RE), pain (PA), social functioning (SF), mental health (MH), vitality (VT), general health (GH)). Two overall scores (Physical Health Summary and Mental Health Summary) are also produced.

Profile measures scores are often transformed for ease of comparison across domains. Taking the SF-36 as an example, each of the 36 items included in the measure are scored on either a 1 to 3, 1 to 4, 1 to 5 or 1 to 6 scale corresponding to frequency or severity level descriptors. Each dimension has between 2 (PA and SF) or 10 (PF) items. To calculate the transformed score, the total raw item score for each dimension is calculated. The raw score for each respondent is then transformed onto the 0 to 100 scale using Equation 6:

((Raw score – min score) / Score range) * 100 (6)

The advantage of profile measures is that they provide detailed information about a patient's QoL from their own perspective. The resulting scores can be used in clinical settings to inform care, and in trials to assess change in QoL over time across a range of domains. However, they can be long, and therefore burdensome for some populations and patient groups. They may also include questions that are not relevant for all patients. For example, a particular domain may not be important, or items investigating a particular severity level may not be required (for

example, if a person has difficulty walking 100 metres, they will also have difficulty walking one kilometre).

A further disadvantage of these measures is that we do not know which items and dimensions are preferred, or considered to be more important than others by patients, as responses are not preference weighted. Therefore, they cannot be used in healthcare decision making which requires the level of preference for QoL outcomes as a result of different interventions to be compared.

2.11. Preference-based measures of health

2.11.1. General structure and principles

In contrast, PBMs generate utilities on a cardinal scale that incorporate preferences and can be used to inform resource allocation decision making. PBMs combine domains of HRQoL into a single preference weighted index score. These instruments typically do not capture the same level of detail as profile instruments, but can be used in the estimation of QALYs.

As described in Chapter 1, PBMs include two elements, a way of describing health or QoL, often referred to as a health state descriptive, or classification, system, and a way of scoring the health states, the utility value set. The descriptive system includes questions that are completed by patients or other populations to reflect their own health and could be generic or condition specific (i.e. specifically developed to measure the impacts of a certain health condition). The value set is based on the preferences of a population for health states described by the classification system. This is often the general public, particularly if the values are to be used in resource allocation in a publicly funded system. A value is assigned to every state that can be described, based on a preference elicitation exercise to produce the overall value set, which is anchored on the full health to dead utility scale.

There are a number of existing HRQoL focused PBMs that are used internationally in healthcare decision making. Many PBMs are designed to be generic so that they can be used to compare outcomes across diverse health conditions and treatments. These include the EQ-5D-3L [77], EQ-5D-5L [78], SF-6D [79, 80] and AQoL [81, 82]. The most widely used of these are the two versions of the EQ-5D ([Scuffham, Whitty [83]] which are accepted by reimbursement agencies internationally [84]

PBMs can be developed *de novo* using qualitative methods and psychometrics, but can also be developed by applying psychometric methods to profile measures (see e.g. [85]). The rationale for using existing generic PBMs is that they are validated in various contexts, and are comparatively easy to use across settings. They also provide a standard measure across programs, patient groups and treatments, and are accepted by many Health Technology Assessment (HTA) agencies in their 'reference case' analyses. A major disadvantage is that the descriptive system may not be relevant or sensitive to the impacts of all conditions, or the positive and negative effects of all treatments.

2.11.2. Who should value quality of life?

Before describing existing PBMs, it is worth considering the question of which populations should value QoL. This is an important area of debate [86] as it can affect the way in which the value sets are used for decision making, and also change the characteristics of the value sets estimated. In taxpayer funded health systems such as Australia and the UK, preferences are elicited from general population samples that are representative in terms of age and gender (and sometimes region). This is because it is argued that as taxpayers, the population should have an influence over how their taxpayer healthcare dollars are spent.

The argument against eliciting general population preferences is that they may not have experience of certain health conditions, and therefore may provide ill-informed values. This means that there is an argument for eliciting preferences from patients or those experiencing the condition. Patients will have better knowledge of the condition and the QoL concepts measured and hence provide more accurate values. However, they may have vested interests in providing values more likely to demonstrate larger benefits, as this could increase the chances of favourable resource allocation decisions. Also, valuation tasks such as TTO and DCE are difficult for the general population, and this may be amplified in certain conditions such as dementia [87].

There is evidence of differences between general population and patient valuations in different physical and psychological health areas. However, there is no consistent pattern, and the generic or condition specific focus of the health state descriptions been valued were also important. For example, Gandhi and colleagues [88] found that patients with heart disease and cancer gave lower overall values than the general population for generic health state descriptions. People with dementia were found to give lower values for dementia related health states than the

general population [87], but in in other cases, such as epilepsy, the values obtained from patients and the general population did not significantly differ [89].

In this thesis, the valuation studies use general population samples. This is because using the general population demonstrates how the methods can inform the development of utility value sets for use in decision making in publicly funded systems such as Australia. It also provides preferences for broad and diverse descriptions of health and QoL from a wide sample with a range of different health experiences, and associated preferences.

2.11.3. Example of a generic PBM - EQ-5D

The most commonly used PBM of HRQoL is the EQ-5D [77]. The EQ-5D descriptive system measures health across five dimensions (mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD) and anxiety/depression (AD), with either three (EQ-5D-3L; none, some/moderate and unable to/extreme) or five (EQ-5D-5L; none, slight, moderate, severe, extreme/unable to) response levels. The EQ-5D-3L and EQ-5D-5L are displayed in **Table 1** and **Table 2**. The EQ-5D-5L was developed to increase the sensitivity of the descriptive system to smaller changes in health, and standardise the wording used across the dimensions ([78]).

Value sets for the EQ-5D have been developed internationally [90]. For the EQ-5D-3L, the most influential value set was developed in the UK using TTO[29]. This value set ranges from 1 (for the best health state) to -0.594 (for the worst health state with severe problems on each dimension described as state 33333) and includes 33% of states valued negatively, so equivalent to states worse than dead. The Australian TTO value set was developed by Viney et al [91] and ranges from 1 to -0.217.

Table 1: The EQ-5D-3L descriptive system

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

Table 2: The EQ-5D-5L descriptive system

Dimension	Level	Description
Mobility	1	**I have no problems in walking about**
	2	**I have slight problems in walking about**
	3	**I have moderate problems in walking about**
	4	**I have severe problems in walking about**
	5	**I am unable to walk about**
Self-Care	1	**I have no problems washing or dressing myself**
	2	**I have slight problems washing and dressing myself**
	3	**I have moderate problems washing and dressing myself**
	4	**I have severe problems washing and dressing myself**
	5	**I am unable to wash or dress myself**
Usual Activities	1	**I have no problems doing my usual activities**
	2	**I have slight problems doing my usual activities**
	3	**I have moderate problems doing my usual activities**
	4	**I have severe problems doing my usual activities**
	5	**I am unable to do my usual activities**
Pain / Discomfort	1	**I have no pain or discomfort**
	2	**I have slight pain or discomfort**
	3	**I have moderate pain or discomfort**
	4	**I have severe pain or discomfort**
	5	**I have extreme pain or discomfort**
Anxiety / Depression	1	**I am not anxious or depressed**
	2	**I am slightly anxious or depressed**
	3	**I am moderately anxious or depressed**
	4	**I am severely anxious or depressed**
	5	**I am extremely anxious or depressed**

For the EQ-5D-5L, the recommended valuation process uses a combination of TTO and DCE. This has been done in the UK [92, 93], and produced a value set ranging from 1 to -0.285 (for the worst state which is described as 55555) with 5% of states valued as worse than dead. In Australia, both TTO and DCE specific value sets have been developed. Norman et al [57] used DCE to produce a value set that ranges from 1 to -0.676 (with approximately 30% valued negatively), and Flattery et al [94] used a combination of TTO and DCE to produce a value set ranging from 1 to -0.366, with approximately 7.5% of states valued as worse than dead. This demonstrates that different descriptive systems and approaches to valuing the health states described leads to value sets with differing characteristics.

2.11.4. Example of a generic PBM - Short Form-6 Dimension (SF-6D)

The SF-6D was developed from the SF-36 [75]. The SF-6D measures HRQoL across six dimensions (physical functioning, role functioning, social functioning, pain, vitality and mental health). Each of these dimensions has a set of responses with between 4 and 6 levels. Therefore, the SF-6D classification system describes 18,000 possible health states (see **Table 3** for the classification system).

The UK value set was developed using SG and ranges from 1 to 0.29 [79] meaning no states are valued as worse than dead. In Australia, the utility value set was developed using DCE with duration which produced values with a range from 1 to -0.363, with 5% of states valued as worse than dead [58]. Other country specific value sets have been developed, including in Spain [96], Japan [97], Brazil [98], Portugal [99], the Netherlands [60], China [100] and Hong Kong [101], and many are accepted by international reimbursement agencies [84]. Valuation studies using ranking [102] and Bayesian modelling methods [103] have also been conducted, resulting in lower values than the UK SG exercise.

Table 3: The SF-6D classification system

Dimension	Level	Description
Physical Functioning	1	**Your health does not limit you in *vigorous activities***
	2	**Your health limits you a little in *vigorous activities***
	3	**Your health limits you a little in *moderate activities***
	4	**Your health limits you a lot in *moderate activities***
	5	**Your health limits you a little in *bathing and dressing***
	6	**Your health limits you a lot in *bathing and dressing***
Role Limitation	1	**You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems**
	2	**You are limited in the kind of work or other activities as a result of your physical health**
	3	**You accomplish less than you would like as a result of emotional problems**
	4	**You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems**
Social Functioning	1	**Your health limits your social activities *none of the time***
	2	**Your health limits your social activities *a little of the time***
	3	**Your health limits your social activities *some of the time***
	4	**Your health limits your social activities *most of the time***
	5	**Your health limits your social activities *all of the time***
Pain	1	**You have *no* pain**
	2	**You have pain but it does not interfere with your normal work (both outside the home and housework)**
	3	**You have pain that interferes with your normal work (both outside the home and housework) *a little bit***
	4	**You have pain that interferes with your normal work (both outside the home and housework) *moderately***
	5	**You have pain that interferes with your normal work (both outside the home and housework) *quite a bit***
	6	**You have pain that interferes with your normal work (both outside the home and housework) *extremely***
Mental Health	1	**You feel tense or downhearted and low *none of the time***
	2	**You feel tense or downhearted and low *a little of the time***
	3	**You feel tense or downhearted and low *some of the time***
	4	**You feel tense or downhearted and low *most of the time***
	5	**You feel tense or downhearted and low *all of the time***
Vitality	1	**You have a lot of energy *all of the time***
	2	**You have a lot of energy *most of the time***
	3	**You have a lot of energy *some of the time***
	4	**You have a lot of energy *a little of the time***
	5	**You have a lot of energy *none of the time***

2.11.5. Example of a generic PBM - Health Utilities Index (HUI 2 and HUI-3)

The Health Utilities Index is a generic measure of HRQoL including two classification systems (the HUI-2 and HUI-3) [104]. The HUI measures were developed in Canada [105], and have been used extensively in Canadian population health surveys. In comparison to the EQ-5D and SF-6D, the HUI measures take what is described as a 'within the skin' approach to measuring HRQoL, and therefore excludes dimensions relating to social functioning and activities. The HUI-2 [106]

was developed as a tool to measure the HRQoL of childhood cancer survivors, and includes seven dimensions defined as sensation, mobility, emotion, cognition, self-care, pain and fertility (for the full description, see Appendix 1). The HUI-2 response levels vary between three and five, and this produces a system describing 24,000 health states. The HUI-3 (see also Appendix 1) was developed for administration in the general population, and the content differs from that of the HUI-2. At the overall level, the system includes eight dimensions defined as vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain.

Comparing the two classification systems demonstrates that three dimensions (emotion, cognition and pain) are included on both, but the content of the description differs. Emotion on the HUI-2 is described as distress and anxiety in comparison to HUI-3 which focuses on happiness and unhappiness. Pain on the HUI-2 focuses on severity of pain whereas HUI-3 is concerned with both the severity and impact. Cognition also changes from a focus on learning on HUI-2 to thinking, memory and problem-solving in HUI-3. Regarding the dimensions that differ, HUI-2 sensation was split into three dimensions (vision, hearing, and speech) on HUI-3 which allows for more focused measurement of these issues. HUI-2 mobility is linked to HUI-3 ambulation and dexterity, and HUI-3 does not include self-care or fertility. The HUI utility scoring system was developed using single- and multi-attribute utility functions using data from two preference surveys including VAS and SG. The range of scores is from 1 to -0.03 for HUI-2, and from 1 to -0.36 for HUI3 [107].

2.11.6. Example of a generic PBM - Assessment of Quality Of Life (AQoL)

The Assessment of Quality of Life (AQoL) is a series of measures of HRQoL developed in Australia by Richardson et al [59] that include the AQoL-4D, AQoL-6D, AQoL-7D and AQoL-8D. The AQoL team defined health, as “a state of optimum physical, mental and social wellbeing and not merely the absence of disease or infirmity” [108] which builds on the WHO definition described in Section 2.6.

The AQoL-8D [81] extends the previous instruments, and is therefore the most comprehensive of the set. The AQoL-8D questionnaire includes 35 items that are mapped onto the eight dimensions replicated in Appendix 2. In comparison to the HUI, the AQoL-8D includes both ‘within the skin’ and wider psychosocial issues, with the aim to improve measurement sensitivity in these areas. The eight dimensions are defined as independent living, pain, senses, mental health, happiness, coping, relationships and self-worth. These also map to what Richardson and

colleagues [81] describe as physical and psychosocial 'super dimensions' respectively. The AQoL-8D value set was generated by collecting VAS and TTO data from the general population and mental health patients. Modelling generated scores for each of the eight dimensions that were combined to form final AQoL-8D utilities.

2.11.7. Multiplicative and additive value set modelling

There are a number of ways to model value sets, and the two key approaches can be defined as additive (used for EQ-5D and SF-6D) and multiplicative (used for HUI and AQoL). Additive functions assume that the disutility of level changes within dimensions are not impacted by level changes in other dimensions. Therefore utilities are generally calculated by summing the decrements across the dimensions, although some interaction terms (for example between dimension levels) have been included in the value set calculations. Multiplicative functions focus on assessing interactions between health state levels. This is done to capture the important preference interactions among the levels. In the modelling of the DCE data reported in this thesis, additive models are generally used, with some testing of interactions of differing severity levels.

2.11.8. Limitations of HRQoL focused descriptive systems

Generic PBMs of HRQoL have limitations linked to the descriptive systems and value sets and these can limit their usefulness in healthcare decision making. Regarding descriptive systems, the methods used to develop the dimensions differ, and the processes used can impact the validity of the final descriptive system. By their nature, PBMs cannot include a large number of dimensions, and this restriction means that they are limited in what HRQoL factors they can measure. There is evidence from systematic reviews and both qualitative and quantitative research that suggests that the generic PBMs are not sensitive to the impacts that some conditions have on QoL. Therefore, the psychometric performance differs across conditions with different impacts on HRQoL. Systematic reviews have found some evidence generic PBMs have a number of limitations for use in conditions such as skin problems, vision and hearing, some cancers, and more severe mental health problems such as schizophrenia [109-112]. This includes inconclusive evidence regarding the construct validity of the PBMs, and their sensitivity to change over time. Qualitative work with people with mental health problems, has found that the dimensions included in generic systems may not be sensitive to all of the impacts of the conditions the individual experiences [112].

Analysis using psychometric methods has also tested generic descriptive systems across a wide range of conditions and found that the measures are valid for use in a number of long-term conditions including diabetes [113] and rheumatoid arthritis [114]. However, psychometric validity was lower in a number of health areas including some types of cancer [115] and more severe mental health issues [111]. This is because the main focus is on HRQoL rather than the other QoL impacts that a condition and a treatment may have, for example on social care related quality of life (SCRQoL), social functioning, capabilities or wellbeing. The limits on what is measured may limit the wider applicability of the resulting QALY estimates, and mean that the effectiveness of treatments improving wider QoL areas may be underestimated. It could be argued that these aspects should be included in PBMs used in decision making alongside HRQoL.

There have been calls in the literature to extend the health-related QALY ([116, 117] to include wider QoL concepts, and PBMs have been developed to measure a wide range of perspectives beyond HRQoL, including SCRQoL and capabilities. The availability of multiple measures requires comparisons to understand in which settings, conditions and populations different instruments should be used, and also understand the relationship between them. However, there is limited work comparing these measures and approaches from both a measurement and valuation perspective. Systematic investigation is required to understand the performance of the measures. It is important to understand what the instruments are measuring, and also the relationship between diverse instruments. This also leads to the question of whether measures assessing different concepts can be combined, or broadened, in some way to generate a more flexible way of measuring the impacts of health conditions. These questions will be investigated in the empirical study assessing measurement issues included in this thesis.

2.11.9. Limitations of HRQoL focused value sets

Regarding the development of value sets, the valuation method chosen, and the protocol used are influential in the characteristics of the value set produced. However, it is unclear which method should be used to elicit and understand population preferences. Therefore, there is the need to understand the methods used in more detail, paying particular attention to how the methods could be applied to the valuation of new descriptive systems or measures, to produce value sets that reflect the preferences of the population. If PBMs measuring different concepts can be combined in some way, it is important to understand whether they can be valued on the same scale to produce QALY estimates, and used in economic evaluation. Methodological

investigations can test the valuation of a combination of measures from multiple QoL perspectives, and this work is conducted as part of this thesis.

2.12. Issues with the health-related QALY framework

Although QALYs are widely accepted as a measure of outcomes for use in HTA, they are not without criticism from a conceptual and empirical perspective, and much has been written on the topic [118-120].

Issues with the QALY that are of most relevance to this thesis relate to how the quality component of the QALY is measured and valued, and the subsequent accuracy and relevance of the values produced. As previously described, various generic measures have been developed for use in measuring outcomes to inform the QALY, with a particular focus on the measurement of HRQoL. However, as recently stated in a summary of QALYs by Neumann and Cohen [118] "such generic scales do not always adequately capture a condition's salient attributes (e.g. symptoms of mental illness)". QALYs estimated using traditional measures with a HRQoL focus also do not take wider non health-related benefits into account [119]. It is also unclear which sets of values are most reflective of an individual's actual utility. There is also debate about which populations should assess QoL for use in the QALY (discussed in Section 2.11.2), and the applicability of data from wider populations to inform resource allocation decision making [121]. In addition, QALYs have been criticised by Nobel Laureate Daniel Kahneman [122] who has argued that people do not act according to the axioms of utility theory, adaptation to health states is important, and by not taking account of the various issues we need to "consider the possibility that the utility used in developing the QALY may be wrong".

Wider questions that are the subject of debate include the lack of discrimination across QALYs linked to different condition severities and different trajectories and whether this is fair and equitable to the entire population. For example, should QALYs be equal across all conditions, or should certain conditions or populations, or condition severities be judged differently? The decision making process which makes choices by comparing values across very diverse populations, data sources, and time periods has also been criticised.

2.13. Moving beyond the health-focused QALY

2.13.1. Why is moving beyond the health-focused QALY important?

Interventions can also result in changes in non-health-related QoL (e.g. social care). However, these effects are at best only partially captured by HRQoL focused instruments. Therefore, the health focus of many of these instruments may mean that the effectiveness of treatments improving other outcomes and areas of QoL may be underestimated. There has been recent debate regarding the focus of the quality weight of the QALY in terms of the dimensions included and the aspects of QoL that are measured. Measures to assess broader QoL concepts have been developed, and a selection of these are described below.

2.13.2. Adult Social Care Outcomes Toolkit

The Adult Social Care Outcomes Toolkit (ASCOT; [123]) was developed to measure SCRQoL across eight dimensions (measured by nine questions) with four response levels for each dimension (see **Table 4**). The eight dimensions are defined as control (CO), personal cleanliness and comfort (CL), food and drink (FD), personal safety (SA), social participation and involvement (SP), occupation (OC), accommodation and involvement (AC), and dignity (DI; two questions, of which one is used for the utility weight). The ASCOT utility scale that estimates a social care QALY was derived using TTO and Best Worst Scaling (BWS) and ranges between –0.171 and 1, with '0' being dead, '1' being equivalent to the 'ideal' SCRQoL state, and negative states being equivalent to SCRQoL states worse than being dead.

Table 4: ASCOT Descriptive System

Dimension	Level	Description
Control	1	I have as much control over daily life as I want
	2	I have adequate control over my daily life
	3	I have some control over my daily life, but not enough
	4	I have no control over my daily life
Personal cleanliness and comfort	1	I feel clean and am able to present myself the way I like
	2	I feel adequately clean and presentable
	3	I feel less than adequately clean or presentable
	4	I don't feel at all clean or presentable
Food and drink	1	I get all the food and drink I like when I want
	2	I get adequate food and drink at okay times
	3	I don't always get adequate or timely food and drink
	4	I don't always get adequate or timely food and drink, and I think there is a risk to my health
Personal safety	1	I feel as safe as I want
	2	Generally, I feel adequately safe, but not as safe as I would like
	3	I feel less than adequately safe
	4	I don't feel at all safe
Social participation and involvement	1	I have as much social contact as I want with people I like
	2	I have adequate social contact with people
	3	I have some social contact with people, but not enough
	4	I have little social contact with people and feel socially isolated
Occupation	1	I'm able to spend time as you want, doing things I value or enjoy
	2	I'm able to do enough of the things I value or enjoy with my time
	3	I do some of the things I value or enjoy with your time, but not enough
	4	I don't do anything I value or enjoy with your time
Accommodation	1	My home is as clean and comfortable as I want
	2	My home is adequately clean and comfortable
	3	My home is not quite clean or comfortable enough
	4	My home is not at all clean or comfortable
Dignity 1	1	Having help makes me think and feel better about myself
	2	Having help does not affect the way I think or feel about myself
	3	Having help sometimes undermines the way I think and feel about myself
	4	Having help completely undermines the way I think and feel about myself
Dignity 2	1	The way I'm helped and treated makes me think and feel better about myself
	2	The way I'm helped and treated does not affect the way I think or feel about myself
	3	The way I'm helped and treated sometimes undermines the way I think and feel about myself
	4	The way I'm helped and treated completely undermines the way I think and feel about myself

2.13.3. ICECAP-A

The ICEpop CAPability measure for Adults (ICECAP-A) [124] was developed to measure capabilities on five dimensions (stability, attachment, autonomy, achievement and enjoyment) with four response levels (see **Table 5**). The capability value set is anchored at 1 (full capability) and 0 (no capability), and values can range between 0 and 1.

Table 5: ICECAP descriptive system

Dimension	Level	Description
Stability	1	I am able to feel settled and secure in all areas of my life
	2	I am able to feel settled and secure in many areas of my life
	3	I am able to feel settled and secure in a few areas of my life
	4	I am unable to feel settled and secure in any areas of my life
Attachment	1	I can have a lot of love, friendship and support
	2	I can have quite a lot of love, friendship and support
	3	I can have a little love, friendship and support
	4	I cannot have any love, friendship and support
Autonomy	1	I am able to be completely independent
	2	I am able to be independent in many things
	3	I am able to be independent in a few things
	4	I am unable to be at all independent
Achievement	1	I can achieve and progress in all aspects of my life
	2	I can achieve and progress in many aspects of my life
	3	I can achieve and progress in a few aspects of my life
	4	I cannot achieve and progress in any aspects of my life
Enjoyment	1	I can have a lot of enjoyment and pleasure
	2	I can have quite a lot of enjoyment and pleasure
	3	I can have a little enjoyment and pleasure
	4	I cannot have any enjoyment and pleasure

2.13.4. Limitations of broader PBMs

The ASCOT and ICECAP PBMs also have a number of limitations which may limit their wider use across health and care settings. The narrow coverage of concepts focusing on certain dimensions of capabilities and SCRQoL means that the measures are not sensitive to the integral relationship between health and these concepts. There is also difficulty valuing broader concepts using standard conceptualisations of common tasks such as TTO, where respondents may not be willing to trade life years, or accept a risk of death to avoid the living situations and dimensions described.

2.14. What about combining or broadening existing outcome measures?

There is debate about the use of measures both within a measurement area (for example which

domains of HRQoL should be included, and which instrument is the most acceptable) and in terms of what should be measured (for example should HRQoL, capabilities, SCRQoL or wellbeing be assessed). This presents challenges around whether the measures could be combined, or broadened, and how this could be done. There is limited work available testing the feasibility of broadening different types of outcome measures to produce an instrument that could encompass the range of dimensions that are important in the assessment of the impacts of different health conditions and treatments.

To extend this research area it is important to understand what different instruments are measuring, and how they could be broadened. Broadening measures would allow for a more accurate assessment of the benefits of interventions beyond a narrow HRQoL focus, and psychometric methods can be used for this. It would facilitate comparability in terms of what is measured across multiple settings. This thesis involves an investigation of the potential to broaden measurement in two ways. First extending dimensions within existing PBMs to include other QoL dimensions is tested. Second, analysis investigates whether existing PBM frameworks can be used to provide further information for each dimension whilst also providing preference-based information.

If the measures can be combined or broadened, then the question of whether they can be valued on the same scale for use in economic evaluation arises. This is a key question as it could be argued that preferences used in decision making should be in the context of a broader set of dimensions (so the values for certain dimensions are considered in the context of the set of broader dimensions). In this thesis, the appropriateness of the use of DCEs to value broader measures of health will also be investigated.

The answers to these questions can be brought together to help to understand whether a broader method of measurement and valuation that is useful for decision makers can be developed. This work will add to existing knowledge around extending the measurement and valuation of health and QoL.

2.15. The empirical work reported in the thesis

In the previous chapters, the rationale for investigating the research questions and objectives studied in this thesis have been outlined. In the chapters that follow, a structured review and the three empirical studies are reported. The review focuses on the methods used for DCE for

health state valuation, and this provides the basis for two empirical pieces of work applying DCE to investigate extending the valuation framework. Prior to the valuation work, the study investigating the potential to develop a broader approach to measurement is reported.

3. Using Discrete Choice Experiments to value health states: A structured literature review

3.1. Summary

This chapter builds on the earlier descriptions of DCE in Chapters 1 and 2. It describes a structured review of the use of DCE methods for the purpose of health state valuation. The review investigates the range of approaches used, and highlights related methodological issues. This is done as the methods used for DCE for health state valuation ultimately have an effect on the value sets produced for use in decision making. However, the nature of the different DCE methodological approaches, and the extent of their use, is poorly understood. Section 3.2 reports the methods used for the structured review, Section 3.3 reports the results, and Section 3.4 discusses the key issues raised by the review.

This review has been published. It appears in a leading health economics journal, Pharmacoeconomics [128].

3.1.1. Aims and objectives:

The review is concerned with Aim 1 of the thesis, which is to conduct a structured review of the use of DCEs for health state valuation. The specific objectives of the structured review are threefold:

1. To review the current literature relating to the use of DCE to value generic and condition specific PBMs to establish what approaches have been taken;
2. To provide a detailed summary of the different DCE methods and approaches used for health state valuation;
3. To establish where there are limitations, areas where there is consensus, and understand where further research is required.

The search, paper screening and data extraction process was conducted by the author of this thesis without cross validation from other collaborators or supervisors. Therefore, the review is framed as a structured review of the area rather than a systematic review which requires data to be extracted by multiple researchers.

As outlined in Chapter 2, DCE is only one of a number of valuation methods that could be used to value broader QoL outcomes. This review focused on reviewing DCE literature given that this was the valuation method used in the two empirical preference studies reported in the thesis.

DCE also has methodological issues that are specific to the approach. Focusing on DCE allows for a detailed exploration of these issues, an understanding of the limitations of the approach, and where further research is required.

3.2. Structured review methods

3.2.1. Literature search

Published literature using DCE methods to generate values for PBMs was identified using PubMed (up to 31/05/2018). PubMed was used as it is a comprehensive source of literature in the health state valuation research area. The search terms were developed by the author and supervisory team considering the key words used to describe DCEs and the process of health state valuation (e.g. DCE or conjoint analysis) and PBMs (e.g. multi-attribute utility instruments, EQ-5D or SF-6D). The full search terms are included in Appendix 3.

The titles and abstracts of papers were initially screened, with non-relevant papers and papers not meeting the inclusion criteria excluded at this stage. This resulted in a set of full text papers which were assessed for relevance for the review, and against the inclusion criteria. The set of papers included were categorised as primary data collection or secondary research. Primary studies included those collecting data to derive a value set or test a methodological issue relating to DCE for health state valuation. Secondary studies included those conducting further modelling work on existing primary data collected for another purpose.

3.2.2. Inclusion and exclusion criteria

Papers were included if they used DCE to develop or directly inform the production of value sets for generic or condition specific PBMs. Papers could include any sample (for example the general population or patient groups). Only studies published in English were included. Studies that aimed to develop a value set or to test and compare DCE based methods were included.

Papers were excluded if they used Case 1 or Case 2 BWS methods (see Cheung et al [126] for a review of the use of these methods in health). This was because these methods have different methodological questions to DCEs that compare multiple health states. Case 3 BWS studies were included as these require the presentation of multiple, rather than single, profiles and therefore require methods more in line with DCE based approaches. Choice based methods that valued partial health states, or valued states not derived from a PBM, were excluded as these studies were not designed to develop a value set.

Qualitative studies were also excluded as the focus of this review was to understand the methodological issues underlying the study designs used for the development of value sets. Qualitative work in a health state valuation framework presents health states from a small subset of choice sets, and cannot be used to estimate value sets [56, 130]. Therefore, this work was excluded from the review.

Further exclusion criteria related to papers where DCE data was collected, but the model results were not reported, and the study design procedure was described elsewhere (for example EQ-5D-5L valuation studies using the standardised protocol [47] but where the DCE data that was collected was not used in the estimation of the published value set [128]). In several cases the same study was published in multiple papers (for example in a published report and a journal article). In these cases, the content of both manuscripts was assessed for extra analyses and both were included if additional relevant information was provided.

3.2.3. Assessing paper content and quality

Assessment of the quality of reporting in the papers was guided by the recently developed Checklist for Reporting Valuation Studies (CREATE) [129] (see Appendix 4). This checklist includes 21 items assessing the key elements that should be reported in health state valuation studies across seven headings. These are defined as descriptive system, health states valued, sampling, preference data collection, study sample, modelling, and scoring algorithm. The items are scored either yes or no. The checklist was used as a guide to the appropriateness of the content only, as many of the studies were methodological and were therefore not aiming to directly develop a value set, and so a number of the items were not applicable. This meant that the two scoring algorithm items ("Criteria for selecting the preferred model are stated" and "The scoring algorithm is presented") were only used for the papers classified as value set development, and not for methodological papers (where the 19 items not focusing on the value set algorithm were used). The percentage checklist score of each paper was calculated to allow for comparability.

3.2.4. Data extraction process

Information was extracted from the papers to allow for an assessment of the current use of DCE for health state valuation. Extraction was conducted using a framework adapted from tables used in previous review work, including that of the author [109]. Data was extracted under the following six headings:

1. General study information: Basic information about the study including country, respondent group, PBM(s) included, general methods, aims, and study categorisation.
2. Task and study design: This included task design features such as the inclusion of duration and death, the number of scenarios (for example pairs or triplets), and the number of attributes. Procedural information such as sample size and drop out, number of choice sets overall and per respondent, observations per choice set, the administration mode, and other relevant information was also extracted.
3. Type of designed experiment: This included type of design and health state selection method used, and other relevant information.
4. Modelling and analysis methods: This included data analysis conducted, information about modelling approach used including the functional form and estimation procedures, and other relevant information.
5. Results: This included the main results reported, for example model consistency (in terms of coefficient ordering), value set range and dimension order, and other information.
6. Discussion: This included the main author conclusions, study limitations (from both the authors of the paper, and the thesis), research recommendations, and any relevant further comments.

In the results below, extraction categories one to four are described in detail. Category five is not included in the results reported. This is because the studies had different aims meaning that it is not appropriate to undertake comparisons of the results or directly relate all the methodological and modelling differences across the studies to variation in the approaches used. Category six is used to inform the summary of the current status of DCE for health state valuation, and areas for further research, that are described in the discussion of this chapter.

3.3. Results

3.3.1. Studies identified

Appendix 5 includes a flow chart of the overall literature searching and identification process. The search identified 1,132 unique records, from which 1,052 were excluded at the title and abstract screening stage resulting in the identification of 80 papers for potential inclusion in the review. Following assessment of the full articles, a further 17 were excluded as they did not fit the inclusion criteria, were not relevant to the review, or summarised data published elsewhere. This resulted in 63 papers being included in the review [43-46, 48-60, 92, 93, 130-173].

3.3.2. Findings – General study information

Table 6 lists the 63 papers in chronological order by year of publication and includes study categorisation, and other key characteristics. The majority of papers (55) were primary studies, with 26 categorised as methodological (including testing methodological issues such as anchoring and comparing methods), 19 were categorised as value set development (applying methods to develop value sets) and eight were categorised as both. The nine secondary studies included further analytical work on a primary dataset (n=7), value set development (n=1) and methodological comparisons (n=1). The year of publication of the papers identified demonstrated increased work in this area, with five papers published up to and including 2010, 22 between 2011 and 2015, and 36 between 2016 and the end of the review search period (May 2018).

Overall, 50 of the 78 study populations were from majority English speaking countries including the UK/England (21), the US (11), Australia (11), Canada (5), and Trinidad and Tobago (2). Other countries with multiple studies included the Netherlands (13), Spain (4), Germany (3), Japan (2), Indonesia (2) and France (2), with China, South Korea, Sweden and Thailand all providing one. The majority of samples (55) were taken from the general population. Other populations (including patients, veteran and specially targeted groups) and students were represented in eight and four papers respectively. The main focus was on the EQ-5D, with EQ-5D-3L appearing in 12 papers, and EQ-5D-5L in 29. Other generic health measures valued included the SF-6D (3), EQ-5D-Y (1), CHU-9D (1) and the PROMIS-29 (1). A number of condition specific measures were also valued. These included the cancer specific EORTC QLU-C10D (4 studies), the glaucoma specific GUI (2), the asthma specific AQL-5D (1), and the COPD specific ABC Index (1). Several measures that assess broader issues not specific to a single condition, but which only appeared once, include three instruments generally for older people (the ASCOT, the OPUS, and the ICECAP-Social Care Measure), and instruments assessing carer QoL (CarerQoL-7D; 2 studies), obesity specific QoL (IWQOL-Lite), sexual QoL (SQOL-3D), labour and delivery issues (LADY-X) and child specific behavioural problems (BPI).

Table 6: Study Categorisation

Study	Categorisation			General information			CREATE score (%)
	Year	Primary/secondary	Detail	Country	Population	Measure	
Hakim & Pathak [52]	1999	Primary	Methodological (Comparing)	United States	Veterans	EQ-5D-3L[a]	89.5
Ryan et al [132]	2006	Primary	Value set development; Methodological (Comparing)	United Kingdom	Over 60s	OPUS[b]	81.0
Burr et al [156]	2007	Primary	Value set development	United Kingdom	Patients	GUI[c]	90.5
Ratcliffe et al [135]	2009	Primary	Value set development; Methodological (Testing and comparing)	United Kingdom	General population	SQOL[d]	94.7
Stolk et al [49]	2010	Primary	Methodological (Testing and comparing)	Netherlands	General population; Students	EQ-5D-3L	94.7
Hauber et al [150]	2011	Primary	Value set development	United States	Overweight people	IWQOL-Lite[e]	89.5
Potoglu et al [141]	2011	Primary	Methodological (Comparing)	United Kingdom	General population	ASCOT[f]	94.7
Bansback et al [48]	2012	Primary	Value set development; Methodological (Testing and comparing)	Canada	General population	EQ-5D-3L	95.2
Bailey [157]	2013	Primary	Value set development	Trinidad and Tobago	Students	EQ-5D-3L	90.5
Pullenayegum & Xie [140]	2013	Secondary	Analytical (Anchoring)	Canada; United Kingdom	General population	EQ-5D-5L[g]	89.5
Ramos-Goni et al [136]	2013	Primary	Methodological (Testing)	Spain	General population	EQ-5D-5L	94.7
Norman et al [57]	2013	Primary	Value set development; Methodological (Testing)	Australia	General population	EQ-5D-5L	95.2
Craig et al [158]	2013	Primary	Value set development; Methodological (Testing)	United States	General population	SF-6D[h]	94.7

Bansback et al [55]	2014	Primary	Methodological (Testing)	United Kingdom	General population	EQ-5D-5L	94.7
Mulhern et al [56]	2014	Primary	Methodological (Testing)	United Kingdom	General population	EQ-5D-5L	95.2
Viney et al [50]	2014	Primary	Value set development; Methodological (Testing)	Australia	General population	EQ-5D-3L	95.2
Norman et al [58]	2014	Primary	Value set development	Australia	General population	SF-6D	89.5
Krabbe et al [143]	2014	Primary	Methodological (Testing and comparing)	England; Canada; Netherlands; United States	General population	EQ-5D-5L	100.0
Xie et al [130]	2014	Primary	Methodological (Comparing)	Canada	General population; University staff	EQ-5D-5L	89.5
Gu et al [151]	2014	Secondary	Analytical (Model development)	Australia	General population	EQ-5D-3L	89.5
Van Hoorn et al[159]	2014	Primary	Methodological (Testing)	Netherlands	General population	EQ-5D-3L	94.7
Robinson et al [134]	2014	Primary	Methodological (Testing)	England	Students	EQ-5D-3L	94.7
Hoefman et al [148]	2014	Primary	Value set development	Netherlands	General population	CarerQol-7D[i]	90.5
Craig et al [160]	2014	Primary	Value set development	United States	General population	PROMIS-29[j]	90.5
Scalone et al [131]	2015	Primary	Methodological (Testing)	Netherlands	Students	EQ-5D-3L	94.7
Gartner et al [161]	2015	Primary	Value set development	Netherlands	General population; Women recently given birth	LADY-X[k]	95.2
Rowen et al [53]	2015	Secondary	Analytical (Anchoring)	United Kingdom	General population	AQL-5D[l]	85.7
Hole et al [147]	2016	Secondary	Analytical (Model development)	Australia	General population	EQ-5D-3L	89.5

Mulhern et al [162]	2016	Primary	Methodological (Testing)	United Kingdom	General population	EQ-5D-5L	94.7
Norman et al [163]	2016	Secondary	Analytical (Model development)	Australia	General population	EQ-5D-3L; EQ-5D-5L	94.7
Mulhern et al [59]	2016	Primary	Methodological (Testing)	United Kingdom	General population	EQ-5D-5L	94.7
Shiroiwa et al [164]	2016	Primary	Value set development; Methodological (Testing)	Japan	General population	EQ-5D-5L	85.7
Jonker et al[51]	2016	Primary	Methodological (Testing)	Netherlands	General population	EQ-5D-5L	89.5
Norman et al[165]	2016	Primary	Methodological (Testing)	Australia	General population	QLU-C10D[m]	94.7
Craig et al [166]	2016	Primary	Value set development	United States	General population	EQ-5D-Y[n]	85.7
Craig et al [167]	2016	Primary	Value set development	United States	General population	BPI[o]	85.7
Versteegh et al[168]	2016	Primary	Value set development	Netherlands	General population	EQ-5D-5L	100.0
Bailey et al [169]	2016	Primary	Value set development	Trinidad and Tobago	General population	EQ-5D-3L	90.5
Norman et al[46]	2016	Primary	Methodological (Testing)	France; Germany	General population	QLU-C10D	89.5
Ramos-Goni et al [54]	2017	Primary	Methodological (Testing)	Spain	General population	EQ-5D-5L	90.5
Robinson et al [170]	2017	Primary	Methodological (Testing and comparing)	United Kingdom	General population	EQ-5D-5L	73.7
Xie et al [171]	2017	Secondary	Analytical (Anchoring)	Canada; United Kingdom; Spain; Netherlands; China; Thailand; South Korea; Japan	General population	EQ-5D-5L	73.7
Krucien et al [142]	2017	Secondary	Methodological (Comparing)	United Kingdom	Patients	GUI	89.5

Mulhern et al [45]	2017	Primary	Methodological (Testing)	Australia	General population	EQ-5D-5L	89.5
Goosens et al [152]	2017	Primary	Value set development	Netherlands	General population; Patients	ABC Index[p]	85.7
Huynh et al [146]	2017	Primary	Value set development	United Kingdom	General population	ICECAP-SCM[q]	100.0
Purba et al [139]	2017	Primary	Value set development	Indonesia	General population	EQ-5D-5L	95.2
Hoefman et al [149]	2017	Primary	Value set development	Australia; Germany; Sweden; United Kingdom; United States	General population	CarerQoL-7D	76.2
Jonker et al [44]	2018	Primary	Methodological (testing)	Netherlands	General population	EQ-5D-5L	94.7
Devlin et al [93]	2018	Primary	Value set development	England	General population	EQ-5D-5L	90.5
Feng et al [92]	2018	Secondary	Analytical (Model development)	England	General population	EQ-5D-5L	94.7
Rowen et al [133]	2018	Primary	Value set development	Netherlands	General population	CHU-9D[r]	90.5
King et al [144]	2018	Primary	Value set development	Australia	General population	QLU-C10D	95.2
Cole et al [155]	2018	Primary	Methodological (testing and comparing)	United Kingdom	General population	EQ-5D-5L	94.7
Mulhern et al [43]	2018	Primary	Methodological (testing and comparing)	Australia	General population	EQ-5D-5L	100.0
Purba et al [138]	2018	Primary	Methodological (testing)	Indonesia	General population	EQ-5D-5L	78.9
Gamper et al [153]	2018	Primary	Methodological (testing)	France; Germany	General population	QLU-C10D	94.7
Ramos-Goni et al [137]	2018	Primary	Value set development	Spain	General population	EQ-5D-5L	95.2

Craig et al [172]	2018	Secondary	Value set development	United States	General population	EQ-5D-5L	95.2
Jakubcyck et al [145]	2018	Primary	Methodological (model development and comparison)	United States	General population	EQ-5D-5L	100.0
Craig et al [173]	2018	Primary	Methodological (testing)	US	General population	EQ-5D-5L	94.7
Jonker et al [60]	2018	Primary	Value set development (inc model development)	Netherlands	General population	SF-6D	95.2
Feng et al [154]	2018	Primary	Methodological (testing and comparing)	United Kingdom	General population	EQ-5D-5L	94.7

[a] EQ-5D – Three Level; [b] Older Persons Utility Scale; [c] Glaucoma Utility Index; [d] Sexual Quality of Life Questionnaire; [e] Impact of Weight on Quality of Life-Lite; [f] Adult Social Care Outcomes Toolkit; [g] EQ-5D – Five Level; [h] Short Form – Six Dimension; [i] Carer Quality of Life – Seven Dimension; [j] Patient-Reported Outcomes Measurement Information System – 29; [k] Labour and Delivery Index; [l] Asthma Quality of Life – Five Dimension; [m] Quality of Life Utility Measure-Core 10 dimensions; [n] EQ-5D – Youth; [o] Behavioural Problems Index; [p] Assessment of Burden of COPD Index; [q] ICEpop CAPability – Supportive Care Measure

3.3.3. Findings – Paper content and quality

Table 6 also reports the CREATE checklist score for each paper. The quality of the papers included in the review according to the CREATE checklist was good, with the mean score of 91.6% (range 73.7% to 100%). The CREATE item that was least adhered to was about stating sample size or a power calculation (15 (23.8%) of the studies reported this). This was followed by the items about reporting response rate (47; 74.7%) and goodness-of-fit statistics (49; 77.8%). All papers described the classification system to be valued, and more than 90% described the study design procedures, preference elicitation methods, modelling approaches and the study sample to an acceptable level.

3.3.4. Findings – Choice set and study design

Table 7 reports key characteristics of the choice sets and study designs used. The majority of studies (46) presented DCE choice sets with options described by five or six attributes. Pairs (n=46) and triplets (n= 16), were overwhelmingly used. The use of duration as an attribute in the choice set has become common (n=31) since the first empirical work testing the inclusion of duration was published [6,18]. Alongside this, the inclusion of a third profile (either full health or a dead state) to improve anchoring (n=16) has been tested. The methods used in the studies including duration were diverse. The number of duration levels included ranges from one (i.e. fixed duration across both options) to 27, with the majority (n=16) presenting between three and six. The overall range of actual values used was from two months to 50 years. In the study that used 2 months [80], the range included 21 duration levels with 19 year values from 1y to 20y (excluding 13y), and two month values (2m and 6m). In the study that included 50 years [48] the range included six levels down to one year (1y, 5y, 10y, 15y, 30y, 50y). This demonstrates the level of variation in one key methodological choice. Other methods of anchoring on the utility scale include the use of external data (n=11), mainly from concurrent TTO or BWS studies.

There were also differences in other aspects of the study design. Online data collection was the most common administration mode, particularly in more recent papers (n=37), with face-to-face methods still practiced (n=22). Sample sizes varied widely, as around half of the studies included more than 1,000 respondents. There was also the tendency to include larger pools of choice sets. Overall, 39 studies included more than 151 choice sets. There was also divergence in what the respondents were asked to choose between, with the most common being ‘best’ (n=22), followed by ‘preferred’ (n=16). There were a number of other design features that are widely

employed across the studies including randomisation (at the choice set and dimension level). Dimension level randomisation was shown to have limited impact on the values estimated in three studies [78, 148]

3.3.5. Findings – Type of designed experiment

Table 8 reports the variety of methods used to design the experiment, and software programs used to construct the designs, taking into account that the amount of detail provided by authors varies greatly. Algorithmic approaches were widely reported, and there was the widespread use of both informative (i.e. non-zero) prior values (n=27) and non-informative (zero) prior values (n=16) in different designs. Ten studies used a set of choice sets developed from a starting design by the addition of one, or more, generators. Other design strategies included using full or partial factorials, pivot methods and hand selected choice sets. Ngene (n=14) was the most commonly used software

Many design strategies involved the trade-off between respondent and statistical efficiency. Highly statistically efficient designs generate models with ordered characteristics, but may include choice sets that are difficult for respondents to complete as all dimensions will appear at different levels across the options within each choice set. There was also some evidence that designs using non-zero informative priors are more difficult than those using zero priors [59]. In comparison, choice sets were made more respondent efficient by, for example, introducing a level of within dimension overlap [44].

There were several issues considered in the process of developing designs. First, some studies used a process of blocking choice sets to allow for a range of severities to be included in the set completed by each respondent [3, 77, 135], while in others choice sets from the design were selected at random for each respondent [57, 58, 167]. A small number of studies (n=4) allocated all choice sets to every respondent [4, 8, 140, 158].

Table 7: Study design characteristics

Characteristic	Number of studies or value[a]
Number of health attributes	
4 or less	3
5	24
6	22
7 or more	16
Duration	
No duration	32
Duration	31
Dead or Full Health state	16
Number of duration levels/lifespans	From 1 to 27
Range of duration levels	2 months to 50 years
Anchoring	
No anchoring	18
DCE with duration	31
External data	11
Other (risk, rescaling)	3
Number of options	
Pairs	46
Triplets	16
Quads	1
Sample size	
Range	60 – 8,222
Less than 100 - 500	18
501 – 1,000	13
More than 1,000	32
Choice sets overall	
Range	12 – 3,160
Less than 50	15
51 – 150	11
More than 151	37
Choice sets per respondent	2 – 108
> 10	35
Approximate observations per pair	
Range	7 – 750
Administration mode	
Online	37
Face-to-face	22
Postal	3
Telephone	1
Question asked	
Better/Best	22
Worst	4
Preferred	16
Best and worst	8
Other	9
Unclear	5

[a] Number of occurrences across all studies reported in fields where single number is given (some fields do not add up to 63 as multiple methods were used in a single paper, or the field was not fully reported). Range of values from all papers reported otherwise.

Table 8: Choice set selection methods

Design characteristic	Number of studies
Design type	
D-efficient (zero priors)	16
D-efficient (non-zero priors)	27
Generator developed	10
Full/Fractional factorial	4
Other (Pivoted designs, hand selected)	6
Software	
Ngene	14
SAS	5
Stata	3
Speed	1

Note: Some fields do not add up to 63 as multiple methods were used in a single paper, or the field was not fully reported

3.3.6. Findings – Data analysis and modelling

Table 9 displays the functional form and modelling approaches used in the studies. The majority of studies which used DCE without a duration attribute estimated a main effects only model (n=38), and studies which included a duration as an attribute estimated parameters interacting the attribute levels with duration (n=24). Both approaches produce estimates of the decrement for each level of each dimension. A number of studies moved beyond this approach to produce coefficient estimates for interactions between dimensions, or included extra parameters to estimate the impact of a further decrement on the models (n=20). Examples of this included an extra decrement when any dimension is at the worst level (coefficients sometimes described as the N3 or N5 term) when EQ-5D is the measure being valued [49, 57].

Overall, 17 studies incorporated some modelling of heterogeneity, with the majority using mixed logit methods. Latent Class Logit models were also used. Other models employed included those that estimated the underlying scale of the samples and the potential to pool data using heteroskedastic [174, 175] and Swait and Louviere [71] methods, and hybrid models that combined DCE with other preference data (in particular TTO) [92, 93, 143]. Attribute non-attendance models [155], generalised linear models [5, 84], Generalised Estimating Equations [86] and Fractional polynomials [7, 144] were also employed. More recently, Zermelo Bradley Terry models with functions to take non-linearity of time preferences into account have been developed and implemented [9, 10, 153].

Table 9: Modelling and analysis characteristics

Characteristic	Number of studies
Model functional form	
Main effects	38
Main effects * duration interactions	24
Extra terms and interactions (e.g. N3 and between dimensions)	20
Regression approach	
Conditional Logit (including random effects, stacked, ordered)	30
Probit (including random effects, stacked, ordered)	13
Heterogeneity (e.g. mixed logit, latent class logit)	17
Hybrid models of TTO and DCE	8
Scale assessment models/Poolability	8
ZBT with power function	3

Note: Some fields do not add up to 63 as multiple methods were used in a single paper, or the field was not fully reported

3.4. Discussion

3.4.1. Summary

This chapter reports the first structured review of DCE methods used to develop value sets for resource allocation decision making. The review was conducted in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines where possible (Appendix 6 reports the completed PRISMA checklist). The results demonstrate that a diverse range of study design and analysis methods have been used. Understanding of the relationship between the methodological choices made and the resulting value sets is also limited. Despite this range of methods, the underlying characteristics of the value sets produced are generally consistent with those produced using iterative valuation methods such as TTO. For example, the dimension levels are monotonic (i.e. disutility increases as dimension level severity increases), and values are within a similar range. These results indicate that DCEs are valuable for the development of value sets, but further investigation of the limitations is required. Therefore, they could still be seen as experimental, and users should understand the unique features of the value sets produced using DCE before applying them in healthcare decision making. The increase in publications in recent years indicates that using DCE for health state valuation is a growing research area. The results of this review also suggest some recurring limitations, areas where

there is consensus, and areas where disagreement means that further research is required. Each of these is outlined below.

3.4.2. What are the recurring limitations?

There are a range of limitations with the methods used in this area, some of which are regularly described by the study authors and some of which are implied by the results of the review. The first recurring limitation is around the wider applicability of the samples used to develop population representative value sets. Although the results may not differ across administration modes [86], online samples that opt-in to panel membership and complete surveys for a small reward may differ from the overall population in terms of unobservable characteristics that may affect the resulting values, or the heterogeneity in values. However it should also be noted that other modes of data collection such as face-to-face are also likely to recruit respondents who may not be representative of the broader population, particularly in terms of unobservable characteristics. The second recurring limitation is that it is difficult to measure the quality of online data beyond assessing factors such as the time taken. In face-to-face work it is possible to control the environment and increase engagement, whereas this is less controllable possible in an online setting, where respondents can complete surveys at any time.

There are also methodological limitations around the complexity of the choice sets used. Choice sets that present multiple health dimensions are complex and challenging. This relates to the amount of information respondents have to consider when answering the questions (if they are completed in the way we expect) and this complexity is increased when additional concepts such as duration or risk are introduced [127]. This is a challenge in recruiting samples that are fully representative, as task difficulty may mean that those with a lower level of understanding may be marginalized in decision making processes. Certain populations (for example those with cognitive impairment) will have difficulties completing the tasks. There are also issues with the realism of the options which is compounded by the possibility for potentially implausible combinations of health dimensions and severity levels. These issues may mean that respondents adopt simplifying decision making strategies to complete the tasks.

To counteract some of these issues, researchers have attempted to simplify the choice set is by setting up the design so that only a certain number of dimensions can differ across the profiles within each choice set [77, 152]. As discussed above, this represents a trade-off between the statistical and respondent efficiency, particularly for designs in which only the main effects

model is to be estimated. It is also possible to alter the presentation of the choice set to encourage the respondent to pay attention to the process, although this could result in framing effects. Studies have restricted designs to exclude implausible options, but with many preference-based instruments defining what is and is not implausible involves a degree of researcher judgement, or consistency with previously used implausibility criteria [50, 137].

Another area is the diversity of choices for certain aspects of the methods used, and this makes harmonisation of approaches difficult. An example of this is the duration attribute, where both the number of levels, and the value of the levels, varies widely. This could be problematic given the importance of the duration attribute, not just in respondent choice making processes, but also in modelling precise values.

Fourthly there are some recurring limitations with the analysis that is used. For example, many studies do not model interactions between health dimensions, but it is likely that these are important to respondents when they consider health scenarios, and qualitative work can inform this [130]. There is also debate about the method of anchoring that should be used, and the impact of this on values. For example, should the focus be on anchoring within the choice sets, by including duration, or using data from a different but related preference elicitation task such as TTO.

3.4.3. Where is there consensus?

There are some areas of consensus across studies. For example, notwithstanding the concerns about respondent representativeness and completion, there was general consensus that DCE for health state valuation can be carried out online with general population panels with a level of acceptability. There was also some consensus that it is possible to make the choice sets more amenable to completion and more comprehensible and that doing so will encourage acceptable completion. Models that take heterogeneity into account reflect the diversity of population preferences, but the use of multiple value sets from the same population, or value sets that explicitly model heterogeneity, may have implications for policy and the HTA process. For example, using multiple value sets rather than a single value set to assess different populations or interventions means that direct comparisons are difficult.

3.4.4. What are the remaining questions and what further work is required?

The results of this review also indicate a number of areas where there are questions remaining. First, there are issues around choice question format. This includes the most appropriate choice set (for example whether to include duration as an attribute), and whether to include a third option of immediate death or full health for a shorter duration against which pairs of health states may be compared. It is also clear that the presentation method and wording used impacts on completion [43, 44], and additional work is required to understand how the presentation could be improved further without leading to framing effects and bias. This also leads to questions about the feasibility of combining different QoL concepts in the same DCE framework and format, and this is investigated in one of the empirical studies reported in this thesis.

Second regarding the design of the experiment, further work is required to understand the performance of designs constructed by different strategies, such as generator development and algorithmic construction, and to understand whether the constructed design is improved by the inclusion of prior information obtained from earlier studies in the construction of the design. Finally, the trade-off between statistical and respondent efficiency requires further assessment. This can be informed by a wide ranging comparison of different design methods, and this is also conducted in one of the empirical studies reported in this thesis.

Third regarding modelling, further work is required to establish which interaction effects are important. The most appropriate way to model the data to produce acceptable value sets, and model respondent heterogeneity, also requires empirical comparison

3.4.5. Review limitations

The use of DCEs for health state valuation is an active and ongoing research area, and this review is limited as grey literature such as conference papers and reports were not included. There were also different levels of detail provided across the papers which makes full and comparable data extraction difficult. Another limitation is that it is difficult to compare the validity of the characteristics of the value sets produced for each measure for a number of reasons. Firstly, there is no 'gold standard' value set against which to compare the characteristics of those reported in the paper. For example, most reported value sets have at least some characteristics that could be seen as problematic such as the proportion of states valued as worse than dead [47]. The expected characteristics of value sets (for example the utility range, and proportion of states valued as worse than dead) are not known as there is no gold standard, or revealed preference data, available. Secondly, attributing factors of validity to the methodological issues

reviewed is difficult given other variation that cannot be controlled such as differences in sample demographics and differences in the frequency that different methods are applied. Therefore, comparing within and across measures is difficult, and differences based on methodological changes can only be inferred. This means that validity is difficult to objectively assess and infer based on external criteria.

3.4.6. Conclusions

This structured review provides an up-to-date summary of the methodological features of DCEs for the valuation of health. Given the wide variety of approaches currently used, further research comparing methods would be required before a more harmonised approach could be advocated. The information provided supports those requiring values to make an informed choice about value sets based on DCE. This is important as the methodological diversity means that users should understand the features of the value sets produced before applying them in decision making. This review informs the study design of the two separate pieces of empirical work focused on valuation. The next chapter focuses on the measurement of health and QoL outcomes to further understand the features of the descriptive systems that can be valued using DCE based valuation methods.

4. Assessing the relationship between QoL outcome measures using Item Response Theory methods

4.1. Summary

In moving beyond the health-focused QALY, there are questions relating to the measurement of outcomes in terms of how to describe health and QoL, and what to measure. This chapter assesses these questions. It investigates the relationship between instruments measuring QoL from a using different perspectives. The results are used to draw conclusions about two potential directions for future developments in the measurement of QoL outcomes for use in resource allocation decision making. The first development tested is linked to extending the range of QoL dimensions included in generic HRQoL PBMs. The second development is linked to increasing the amount of information provided within the existing measurement framework of generic PBMs. This could be done by adding further questions linked to each HRQoL dimension, and is described hereon as a 'layered' approach to measurement (defined in Section 4.3). Exploratory Factor analysis (EFA) and IRT are the methods used in this study.

The chapter first introduces and justifies the research. The IRT approach is described in detail given its central role in the analyses reported. An overview of the theory and background of IRT, and the IRT methods available, is provided. The literature using IRT methods for the development of PBMs is then summarised. Following this, the empirical work using EFA to assess the dimensionality of the instruments and items included is reported. This is followed by a description of the empirical work using IRT to examine broadening the measurement of QoL. Finally, a discussion of the key findings, and a summary of how they inform potential developments in the measurement of QoL, is presented.

4.2. Introduction

As described in Section 2.12, the narrow focus of generic PBMs measuring HRQoL may lead to the potential benefits of interventions with a wider impact of QoL being underestimated. The calls in the literature to move beyond the health-related QALY [119] have discussed the descriptive limitations of HRQoL focused instruments [120], and there are concerns that generic instruments do not capture all of the important QoL domains. Alongside this, new PBMs to measure a range of perspectives beyond HRQoL have been developed.

The availability of PBMs taking different perspectives on the measurement of QoL raises questions about what concepts and dimensions can and should be incorporated into measures and utility value sets. These questions are important, as with a range of different instruments available, there are choices to be made about which instruments to use for the basis of decision making. One way to tackle this issue is by investigating how different measures could be used to extend the dimension coverage and broaden the scope of the concepts assessed within existing PBMs. Broadening the concepts of QoL measured in a PBM framework could lead to value sets with applicability and sensitivity across wider health conditions and populations.

One way to test broadening the scope of what is measured is to investigate broadening the QoL dimensions that are included in PBM frameworks. This means developing an understanding of where there is overlap amongst the dimensions measured across instruments, and which dimensions provide wider information, and therefore broaden the coverage of QoL. This can be investigated using existing instruments, and allows for an investigation of whether measures assessing different concepts can be combined to generate a broader measure of the QoL impacts of conditions. This is the first potential extension of the measurement of QoL tested in this study, and is done using EFA methods (described as Extension 1). If a range of concepts are identified, then work to develop value sets could be focused on broader QoL frameworks. The valuation studies conducted later in this thesis demonstrate how this could be conceptualised.

4.3. Defining a 'layered' approach to measurement

As described in Chapter 2, PBMs include dimensions which are measured using a single question, partly because this facilitates the generation of utility values, where concise health state descriptions are required for valuation. However, in this framework, detailed information about a patient's QoL is limited. For example, in the EQ-5D-5L, mobility is measured by problems in walking about, and does not capture broader physical functioning concepts such as the ability to do chores, or vigorous and moderate activities, which could be important.

To address this, it is possible to investigate whether more information about each PBM dimension could be captured, whilst still retaining the ability to derive utilities for use in resource allocation decision making. This could be conceptualised as a 'layered' approach to measurement, and is the second potential extension to the measurement of QoL tested in this study (described as Extension 2), and builds on the dimensionality assessment conducted for Extension 1.

For example, a layered instrument could include a higher level preference-based measure based on the EQ-5D (Layer 1), and a profile score based on a set of items measuring the same underlying domain as the Layer 1 dimension (Layer 2). Such an approach would combine the benefits of a preference-based scoring system with the benefits available from the more detailed information that is elicited from profile measures of QoL, and can be used in, for example, clinical decision making and trial settings.

The feasibility of developing this approach can be investigated by assessing the psychometric relationship between items from a range of instruments that measure overlapping QoL constructs. For example this can be done using IRT analyses, and using this analyses would combine the benefits of PBMs with the measurement precision available using IRT. To fully benefit from such an approach, a set of dimensions enabling the estimation of utility values from Layer 1. The second level of measurement can build from the PBM dimension, and link back to this using IRT based analyses of the overall set of items measuring the domain, which would facilitate IRT based scoring, and the implementation of computer adaptive testing (CAT). However, exploratory development taking a dimension-by-dimension approach to understand the feasibility, and the methodological issues that arise, is warranted. The approach can also be conceptualised, and the feasibility tested, using items from existing measures of QoL. The layering approach provides a different perspective on the broadening of QoL measurement, as it looks to extend the information provided about domains included in existing PBMs with available value sets.

4.4. Why is this research important?

The research conducted in this chapter is important, as broadening outcome measurement to include wider QoL concepts, and increasing the information provided within existing instruments, could improve the evidence available for resource allocation and clinical decision making. It is inevitable that when evaluating interventions, instruments that demonstrate the largest difference for particular intervention and conditions will be preferred. However, this leads to a lack of comparability of utility values. By understanding how different outcome measures relate to each other terms of what they are measuring, and what additional information can be added within QoL dimensions we can calibrate utilities to inform decision making. The comparative work conducted here is a fundamental first step in this process.

4.5. Aims and objectives

The aims of the empirical research reported in this chapter are to assess the relationship between QoL outcomes from a range of existing measures and investigate two extensions to measurement of outcomes for use in economic evaluation. These are:

Extension 1: Extending the range of QoL dimensions included in PBMs

Extension 2: Increasing the amount of information provided by existing PBMs by adding further questions as part of each dimension, and exploring the feasibility of a 'layered' approach to measurement

The instruments included in this study measure broad QoL including HRQoL (EQ-5D-5L, SF-36 and PROMIS-29), SCRQoL (ASCOT), capabilities (ICECAP-A) and wellbeing (WEMWBS). These instruments were chosen to allow for a detailed comparison of the measurement characteristics of and relationship between a broad set of relevant domains.

This empirical work links to two of the overall aims of this thesis. These are Aim 2 (to investigate the relationship between a range of QoL outcome measures assessing different concepts, and the potential for broadening the QoL concepts assessed in an extended measurement framework), and Aim 3 (to investigate the potential for providing further descriptive information within the existing framework and dimension structure of PBMs).

4.6. Item Response Theory

IRT methods (summarised by Fayers and Machin [76] and Edelen & Reeve [176]) are a set of theoretical approaches and associated practical methods used for the construction of measurement instruments [76]. IRT was developed for use in an educational setting but has since gained prominence in the development of patient-reported outcome measures.

4.6.1. What is the IRT model, and what does it estimate?

IRT is an umbrella term for a set of generalised linear models that link observed item responses to respondents' location on an unmeasured underlying latent trait (described as the 'theta' (θ) scale). The theta scale builds on the assumption that a set of items measure a unique and identifiable continuous latent trait, and models an unobservable continuous dimension that is assumed to be unidimensional (i.e. measuring one construct). Applying an IRT to a set of items is said to 'calibrate' those items on a unidimensional theta scale. In the case of patient-reported outcomes theta represents a dimension of health of QoL that is measured by responses to a set

of items assessing the same unidimensional concept. The scale represents a continuous severity range across theta. The most common model is a logistic item response model. The simplest form of this is the one parameter model as displayed in Equation 8:

$$\boldsymbol{P(\theta_j)} = \frac{\boldsymbol{exp\{\theta_j - b\}}}{\boldsymbol{1 + exp\{\theta_j - b\}}} \quad (8)$$

In this model, $\boldsymbol{\theta_j}$ is the latent variable for person **j**, and **b** is an item threshold parameter. Item threshold parameters are estimated from an item characteristic curve. An item characteristic curve is a logistic function that models the probability of responses to an item conditional on the severity of theta, and this is done for each response level. The threshold parameter describes the level of theta necessary to transition between item response levels, and endorse the next level, with a probability of 0.5. For example, for a dichotomous (yes/no) item response, the point of a theta scale where the probability of answering 'No' and 'Yes' is 0.5 provides the threshold parameter.

Item characteristic curves are also estimated for polytomous item responses, and can be used to understand response level ordering. The probability of responding to each level is a function of the increasing response level severity, and the severity of theta. As the threshold parameters represent the transition between levels, the number of threshold parameters produced for each item equates to the number of response levels minus one.

Two parameter logistic IRT models are also widely applied. This model is described in Equation 9, and extends the one parameter model to estimate an item discrimination, or slope, parameter (**α**) which can vary between items.

$$\boldsymbol{P(\theta_j)} = \frac{\boldsymbol{exp\{\alpha(\theta_j - b)\}}}{\boldsymbol{1 + exp\{\alpha(\theta_j - b)\}}} \quad (9)$$

Slope parameters provide a single figure estimate of how particular items discriminate at different levels of theta. The slope $\boldsymbol{\alpha}$ is a function of the threshold parameters (***b***) and theta. Generally speaking, items with larger slope parameters, or steeper slopes, provide more discriminatory information which is provided over a narrower range of the latent trait.

An item information function that displays the information provided by an item across theta is also estimated. These curves indicate the points of theta where the item is providing the most discriminatory information and are therefore useful in selecting items that are sensitive at different points of theta. This is important, as items with less steep slopes, and hence lower

discrimination, may also provide a level of information at different points of theta. These functions can be calculated using Equation 10, which demonstrates that the maximum value for item $\boldsymbol{i}$ is linked to the square of the slope parameter $\boldsymbol{a}$, where a larger $\boldsymbol{a}$ results in more information. Also of note is that the maximum information is obtained at the threshold parameter values.

$$\boldsymbol{a_i^2\{1 + exp[-(\theta - b_i)]\}^{-1}\{1 - exp[-(\theta - b_i)]\}^{-1}} \quad (10)$$

An example of how item characteristic curves and threshold parameters are operationalised is given in **Figure 11**. This displays the characteristic curves for an item asking about whether an individual is depressed on a five level frequency scale from none of the time (curve 0), a little of the time (curve 1), some of the time (curve 2), most of the time (curve 3) and all of the time (curve 4). The theta scale is presented on the X axis and measures depression (from a low level of depression represented by negative numbers, to high levels represented by positive numbers). The Y axis is the probability of responding to each severity level. The Figure demonstrates that at low levels of depression, the probability of responding 'none of the time' is high for the first part of theta. As latent depression increases, respondents become more likely to respond using the remaining response levels. The point where the curves cross translates into the threshold parameters (equating to particular values of theta). For example, the item in **Figure 11** has four threshold parameters at theta values of 0.05, 0.78, 1.56 and 2.24. The probability of responding to each severity level is ordered as theta increases (as each response level curve reaches its peak as theta severity increases). Therefore, the response categories are operating as expected.

Figure 12 maps out the information curve based on the discrimination parameter and item characteristic curves for the same depression item. Again, the X axis represents latent depression, but now Y represents the information provided as a function of the threshold parameters. The characteristics of the information curves can be observed to understand the nature of the item information provided across theta. A low level of information is provided at the less severe range of theta (given the high probability of answering 'none on the time'), but information increases as severity increases.

Figure 11: Example of Item Characteristic Curves

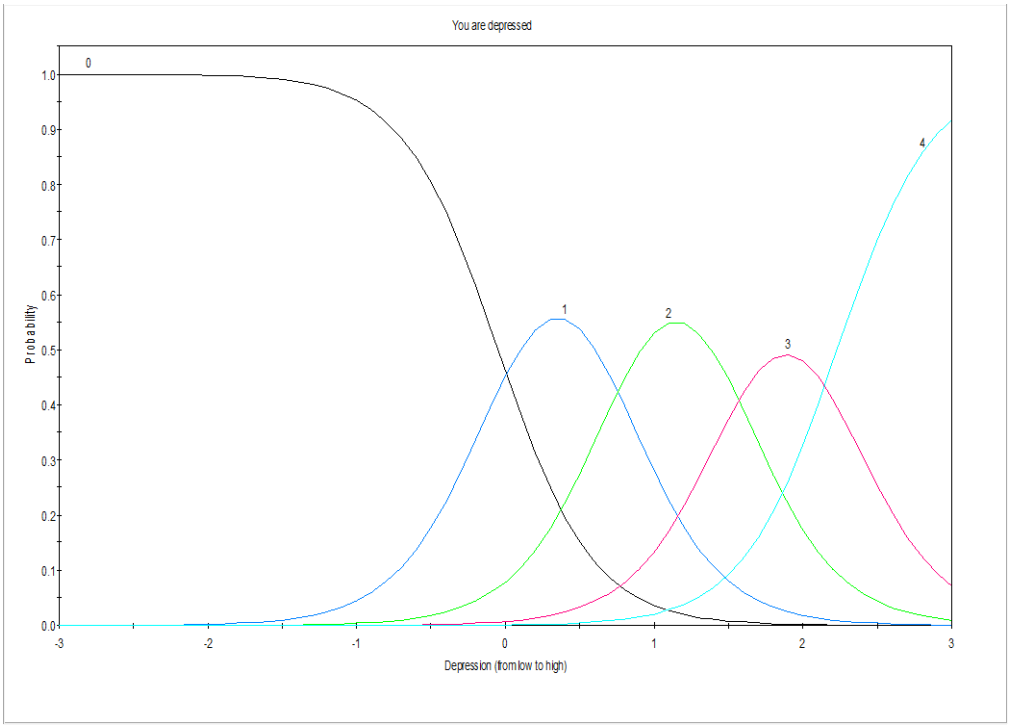
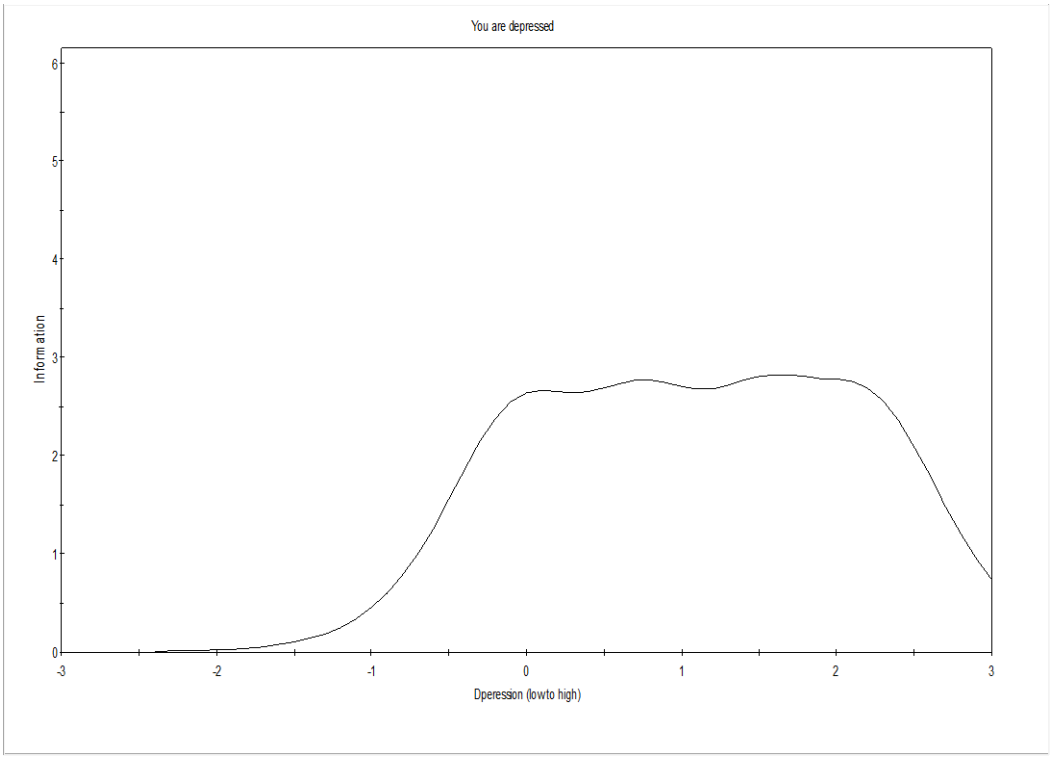


Figure 12: Example of Item Information Curve



4.6.2. What IRT models are available?

Table 10 summarises the most common one and two parameter IRT models. A number of these models have been used for the development and assessment of outcome measures. As discussed in Section 4.4.1, the one parameter IRT model describes the latent trait of the respondents and produces item threshold parameters. The Rasch model [177] is another form of one parameter model. The Rasch model constrains the discrimination parameter to be the same across all items meaning that the models are more parsimonious. However, the estimation of a single parameter, and the fixed discrimination parameter limits the information available for each dimension, as restricting the discrimination parameter may lead to increased evidence of item misfit.

There are a number of two parameter models available to estimate threshold and slope parameters. A commonly used two parameter model in the development of outcome measures is the graded response model [178]. This model is applicable to polytomous response level data with a logical ordering of response levels. This model is applied in the study reported here (for justification, see Section 4.14.1). The model derives the probability of a response for a particular item in a test as a function of theta and the item parameters. The cumulative probability, or the probability of responding in or above a given response category, is modelled. The probability of responding in a specific category is modelled as the difference between two adjacent cumulative probabilities. The graded response model takes the form of Equation 11:

$$\boldsymbol{P\left(X_{ijk} = k|\theta_j, b_{ik}, a_i\right)} = \frac{\mathbf{1}}{\mathbf{1}+e^{-a_i\left(\theta_j - b_{i,k}\right)}} - \frac{\mathbf{1}}{\mathbf{1}+e^{-a_i\left(\theta_j - b_{i,k+1}\right)}} \quad (11)$$

where **i** represents the number of items, **j** the number of people, and **k** the number of response categories, and $\mathbf{X_{ijk}}$ represents response k to item **i** for person **j,** $\mathbf{a_i}$ is the discrimination parameter for item **i**, and $\mathbf{b_{ik}}$ is the thresholds for response level **k** of item ***i***. The thresholds represent the point between response level scores and lead to the graded response model a where $\mathbf{P\ (X_{ijk} = k|\theta_j, b_{ik}, a_i)}$ is the probability of responding at response level **k**.

A three parameter IRT specification can also to extend the model further. The three parameter model builds on the same assumptions as the two parameter model, but introduces a third 'guessing' parameter. This parameter was introduced in educational testing to attempt to model the level of guessing for certain items based on an individual's response profile. However, guessing is a response strategy when the respondent is unsure which answer is correct, and in the completion of outcome measures, the interpretation of this is unclear [176].

Table 10: Summary of IRT models

Model	Response format	Characteristics
One parameter logistic/Rasch model	Dichotomous	Item discrimination is equal, threshold can vary across items
Two parameter logistic	Dichotomous	Both discrimination and threshold parameters can vary
Nominal	Polytomous (no ordering of levels)	Discrimination can vary across items
Graded Response	Polytomous (ordered levels)	Discrimination can vary across items
Partial Credit (Rasch model)	Polytomous (ordered levels)	Discrimination is equal across items
Rating Scale (Rasch model)	Polytomous (ordered levels)	Discrimination is equal across items, and threshold distance is equal
Generalised Partial Credit	Polytomous (ordered levels)	Discrimination can vary across items
Three parameter logistic	Dichotomous	Discrimination and threshold parameters can vary. Non-zero lower asymptote estimated for guessing parameter

4.6.3. Assumptions of the IRT model

IRT models have a number of assumptions that are considered in the analysis process and interpretation of the results. The assumption of monotonicity implies that the probability of responding to an item response level that represents a higher level of theta should increase as the underlying level of theta increases. This is demonstrated by the item characteristic curves shown in **Figure 11**.

Unidimensionality assumes that a single latent construct accounts for the common variances for the items within the model generated. Unidimensional models are applied in this study to examine measurement Extension 2.

Local independence is another key assumption. This states that item responses are independent of each other after controlling for the underlying construct being measured. Examples of local dependence are items that are linked, or very similar within a dimension. Local dependence can lead to parameter estimates that differ in comparison to those reported when items are independent of each other. If dependence is identified, item content can be considered for overlap, and the least informative items can be removed from the models.

4.6.4. How is IRT useful in developing and assessing QoL outcome measures?

IRT informs the development of outcome measures in a number of ways. A key feature of IRT used in the development of outcome measures is ability to describe the measurement sensitivity of items, or sets of items, within a dimension. Item thresholds can be used to highlight the ranges of theta where particular items are more sensitive. The information provided by each item within an overall dimension, and the information provided by all items in a particular dimension, is also informs the development process. By understanding item discrimination, measures of the same underlying construct, but formed of different numbers of items, can be calibrated. This means that it is possible to score individuals using theta, which enables direct comparisons on a common metric [179, 180].

Item performance indicators can be estimated to understand the validity of items within dimensions. These include differential item functioning (DIF), which assesses the extent to which the meaning of, and therefore responses to, items are the same across demographic groups. The models also assess item fit to underlying unidimensional models. The ordering of item responses such that the probability that the expected severity level is endorsed at different points of theta is also tested. Finally, items with conceptual overlap can be highlighted by assessing outliers to the local independence assumption. These indicators are used to develop instruments with strong measurement properties and precision.

4.6.5. Overview of literature using IRT methods in PBM and HRQoL measure development

Extensive work has used IRT to develop outcome measures, and there are a number of overarching areas of work. For example, IRT methods have been applied to develop and test generic and condition specific fixed form HRQoL measures, as well as item banks for CAT administration, that demonstrate a high level of measurement precision. They have also been used to develop PBM descriptive systems, and, to a lesser extent, compare the performance of PBMs. There are also published comparisons of different IRT and classical psychometric approaches to the assessment of outcomes. Given the extent of this work, an overview of this published literature is provided here.

A broad area where IRT methods have been applied is in the development of fixed form profile measures. Both one parameter Rasch and two parameter graded response IRT models have been used to generate mental health instruments with measurement sensitivity [181]. Two parameter IRT models have been applied to QoL data to develop new instruments for other

health conditions. Watt et al. [182] used the graded response model to select items for an instrument assessing thyroid related QoL concerns. A shortened version of the 23 item breast cancer specific version of the EORTC with 12 items was developed using the same modelling approach [183], with items excluded based on the level of discrimination. Graded response methods have also been applied to existing sets of items to develop and assess refined instruments to assess the impacts of low vision [184, 185], disability related to multiple sclerosis [186] and stroke [187] amongst others. The common theme of this work is applying IRT to newly developed or existing items to assess item performance, and generate more precise instruments.

IRT has proved advantageous in the development of calibrated item banks due to the ability to calibrate items and respondents on an underlying metric. Calibrated item banks are the basis of CAT which have a high level of measurement precision. Bjorner et al. [188] and Petersen et al. [189] analysed data from European cancer studies using the generalised partial credit model to select items from the emotional functioning and fatigue domains of the EORTC QLQ-C30 and develop CATs which improved measurement precision. The most well-known CAT based instruments are the PROMIS item banks which are generic, rather than disease specific, and were developed using a graded response model. Large item pools were assessed, with poorly performing items removed. The remaining items were calibrated to enable CAT [190] which builds on the basis that each item's psychometric characteristics are known. The flexible system can therefore iteratively deliver a targeted set of items from the underlying calibrated item bank to patients based on his or her previous item responses, with all patients scored on the same theta scale. The generic nature of the item banks allows for comparisons across populations and conditions

IRT methods have also been used to develop PBMs, with a focus on Rasch analysis (see [191] for a review). Rasch has been used to develop both generic and condition specific PBMs [85], and measures are available in a range of conditions including cancer [144, 192], dementia [193] and epilepsy [90] amongst others. The process of developing a health state classification system involves adapting an existing measure of HRQoL into a shorter version that is amenable to valuation using psychometric analysis to understand the dimensionality of the instrument. Items from the profile measure are tested for validity, and selected to represent the PBM dimensions [177]. In more recent developments, a two parameter IRT model has also been used in the development of a generic mental health instrument [194] A pool of items was generated using

qualitative methods, and IRT was used to reduce this pool to a smaller more precise instrument by assessing the information provided by each with dimensions.

IRT has also been used to assess the performance of PBMs. Izumi et al [195] compared the EQ-5D-3L, EQ-5D-5L and HUI3 in stroke patients using a two parameter model and found good discrimination and wide threshold coverage for all instruments. Fryback et al [196] used IRT to understand the relationship between the EQ-5D, HUI2, HUI3, and SF-6D. They found limited evidence of a relationship across the instruments, suggesting that diverse QoL concepts are measured.

There have also been a number of comparisons of IRT methods. For example, Petrillo and colleagues [197] compared two parameter IRT, Rasch and classical test theory (CTT) using a QoL measure assessing vision problems. They found that IRT and Rasch provided extensive information regarding scale improvements, and highlighted poor fit. CTT was useful for identifying items that were redundant. Stover et al [198], Nolte et al [199] and Cleanthous et al [200] (summarised by Bjorner [201]) used data on the PROMIS depression item bank to also compare two parameter IRT, CTT and Rasch respectively and found differing agreement across model indicators. It was found that CTT and IRT gave similar indicators of discrimination. IRT and Rasch gave similar item thresholds which are not estimated in CTT. Agreement about which items displayed local dependence was low across all three methods. IRT and Rasch consistently identified items that displayed evidence of DIF (which is not estimated in CTT). Measurement precision was similar between IRT and Rasch. Capellieri et al [202] review the use of CTT and IRT for evaluating outcome measures, and conclude that both CTT and IRT are valid to support the maximisation of the content validity of different instruments.

4.7. Description of empirical study

4.7.1. Data and study design

The study involved the collection and analysis of data across a range of instruments. These were implemented through an online survey administered in a sample of the Australian general population, and a further group of respondents self-reporting commonly occurring health conditions (diabetes, mild to moderate depression or anxiety, general pain and arthritis). The measures included in the survey are summarised in **Table 11**. They can broadly be defined as preference-based generic HRQoL, non preference-based generic HRQoL, preference-based

wider QoL (including social care and capabilities), and wellbeing.

The HRQoL focused PBMs included in the study were the EQ-5D-5L [78] and the SF-6D [79, 80]. These measures were chosen as they are the most widely used generic PBMs with different conceptualisations of HRQoL. Therefore, they have both convergence and divergence in terms of what is measured. The HRQoL profile measures were the SF-36v2 [203] (from which the SF-6D is estimated) and the PROMIS-29 [204]. These were chosen as they assess a wide range of health concepts across multiple item dimensions, so are useful to assess overlap with wider measures of QoL. The non health-focused PBMs were the ASCOT measure of SCRQoL, and the ICECAP-A measure of capabilities (see Sections 2.13.2 and 2.13.3). These were chosen to provide a comparative and complementary perspective to the HRQoL instruments. The wellbeing measures included were the WEMWBS [205] and ONS-4 [206]. These were included to understand how wellbeing intersects with measures of wider QoL.

A broad range of instruments were included in the survey to allow for comparisons across different measurement perspectives at both the overall and item level. Demographic information, questions about general health status and co-morbid health condition data were also collected. Appendix 7 includes the full content of the survey, and the instruments included.

The questionnaires were placed into blocks based on the measurement perspective taken, and the content of the questions. Block 1 included the EQ-5D-5L, Block 2 included the non health-focused PBMs (ASCOT and ICECAP), Block 3 included the HRQoL profile measures (SF-36 and PROMIS-29), and Block 4 included wellbeing measures (WEMWBS and ONS-4). The order of the blocks was randomised to generate 24 survey versions including all possible Block orders. Respondents were randomly allocated to one of the versions. This was done to avoid the impact of response fatigue (and therefore potentially less precise data from measures appearing at the end of the survey that might be a concern if all respondents completed them in the same order). The feasibility of collecting outcomes data online in Australia has been demonstrated by the Multi-Instrument Comparison (MIC) study [207], and this study adapted and extended the MIC approach to include measures with a broader QoL and wellbeing focus. Generic profile measures that were not included in that study (the PROMIS-29) were also administered.

Table 11: Measures included in the Health Measurement Study

Measure	Description
Health-related quality of life (preference-based)	
EQ-5D-5L	EQ-5D-5L [78] assesses HRQoL on five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) with five response levels (none, slight, moderate, severe, extreme/unable to).
SF-6D	SF-6D [80] preference-based score is derived from completions of the SF-36. SF-6D assesses HRQoL on six dimensions (physical functioning, role functioning, social functioning, pain, vitality and mental health) with a variety of response levels (from four to six).
Wider QoL (preference-based)	
ASCOT	ASCOT [126] measures social care related QoL on eight dimensions (control over daily life, cleanliness and comfort, food and drink, personal safety, social participation and involvement, occupation, accommodation, and dignity).
ICECAP-A	ICECAP-A ([127] measures capabilities across five dimensions (attachment (ability to have love, friendship and support), stability (ability to feel settled and secure), achievement (ability to achieve and progress in life), enjoyment (an ability to experience enjoyment and pleasure) and autonomy (an ability to be independent)).
Health-Related Quality of Life (non preference-based)	
SF-36v2	SF-36 [75] is a measure of HRQoL producing eight dimension scores from 36 items (physical functioning, role physical, role emotional, pain, social functioning, mental health, vitality, general health). Two overall scores (Physical Health Summary and Mental Health Summary) are also produced (but are not used in this thesis).
PROMIS-29	PROMIS-29 [204] measures health on seven dimensions (Depression, Anxiety, Physical Functioning, Pain Interference, Fatigue, Sleep Disturbance, and Ability to Participate in Social Roles and Activities) with four items per dimension, and an additional pain intensity item.
Wellbeing (non preference-based)	
WEMWBS	WEMWBS [205] is scale of 14 items that are positively worded (given the focus on assessing population wellbeing) with five response categories.
ONS-4	Questions used in international household surveys assessing wellbeing of four questions with ten-point response scales (life satisfaction, health satisfaction, feelings that life is worthwhile, happiness, anxiety).

4.7.2. Respondents and recruitment

Both general population and patient group respondents were recruited for inclusion in the study. All respondents were members of opt-in research panels managed by the panel company

Survey Sampling International. The patient groups included respondents reporting a range of physical and mental health conditions. The conditions were mild to moderate depression and anxiety, diabetes, arthritis and pain, and were selected as they have a range of impacts on different areas of QoL. The survey also asked a detailed multiple choice question about whether they had a long-term health condition. This was done as past online surveys [48, 55] have found that a substantial number of general population panel members report having a long-term health condition.

Respondents were invited by the panel company to take part via e-mail or website advertising, and clicked a link to access the survey. They then read information about the project including the general aims, survey content, and information about data security and confidentiality. They were also informed that they could stop at any time. Following the information page, consent to take part was required. Respondents completing the full survey were provided with a small incentive for taking part if they fully completed the survey in more than the minimum time of four minutes. The amount of the incentive was based on the policies of the online panel they were recruited from. Four minutes was used to exclude very fast completers whilst also being in line with other past online surveys [55], to avoid potential loss of response quality. The study was ethically approved by the Centre for Health Economics Research and Evaluation, University of Technology Sydney, program ethics process (application ref: UTS HREC 2015000135).

4.8. Descriptive assessment of the sample and outcome measures

Descriptive assessment of the sample and responses to the measures was conducted. This was done to understand item response patterns. The analyses conducted are described below.

4.8.1. Sample demographics and survey completion process

The overall number of respondents accessing the survey, dropping out at various stages of the survey, and fully completing all of the questions was assessed. The demographic characteristics and self-reported health of the sample who fully completed the survey were descriptively assessed. As various patient groups were included alongside the general population sample, the aim was not to recruit a sample representative of the Australian population in age and gender.

4.8.2. Scoring the measures

Descriptive analysis of responses was conducted at item and dimension level. At the dimension level, each measure was scored according to standard scoring approaches.

4.8.3. Scoring the PBMs of HRQoL

The EQ-5D-5L was scored using values based on the Australian population. The value set used was the 'Australian Pilot' [57] which was developed using a DCE with duration administered online to a representative sample of the Australian population. It has a range from 1 (for the best health state described) to -0.676 (for the worst health state described). For the SF-6D, the Australian value set that has range of 1 to -0.363 [58] was used.

4.8.4. Scoring the PBMs of wider QoL

The ASCOT was scored using the UK value set [126] which was used as no Australian values are available. The ASCOT value set was developed using TTO and BWS, and resulted in a range of 1 to -0.171. For the ICECAP-A the UK value set [127] was also used as there are no Australian values available. This was developed using BWS and ranges from 1 (full capability) to 0 (no capability).

4.8.5. Scoring the profile based HRQoL instruments – SF-36

The eight SF-36 dimensions were scored using the standard process described in the SF-36v2 scoring manual [203], and replicated in Section 2.10. The transformed scores range from 0 to 100 for each dimension, where a high score is indicative of a better level of HRQoL or functioning.

4.8.6. Scoring the profile based HRQoL instruments – PROMIS-29

The PROMIS-29 was scored at the dimension level using the IRT based theta scores [204]. PROMIS-29 includes 28 items that have five severity levels scored from one to five, and seven dimensions which each include four items. The instrument also includes a single pain intensity item scored on a 0 to 10 scale. To score each dimension, a raw score was calculated by summing the score from each of the items in each dimension (giving a range of possible raw scores from 4 to 20). Each of the possible raw scores was then linked to an IRT based theta score, a standardised score based on the United States population with a mean of 50 and a standard deviation of 10. This scoring means, for example, that a score of 60 is one SD better than the average of the United States general population. The raw scores calculated for the sample were linked to the theta scores using the conversion tables supplied by the PROMIS research group [208].

Whether a higher or lower PROMIS T score reflects better or worse levels of the construct being measured depends on the direction of the wording of the item response levels. For the negatively framed dimensions (anxiety, depression, fatigue, sleep, and pain interference), a higher score is indicative of higher levels of the concept measured. For the positively framed dimensions (physical function, and ability to participate in social roles and activities) a higher theta score represents a better level of functioning.

4.8.7. Scoring the WEMWBS

The WEMWBS produces 14 item scores with five levels corresponding to poor (level 1) and good (level 5) wellbeing. The aggregate total of the scores for each respondent was calculated to give an overall wellbeing indicator. Scores range from 14 to 70, with high scores indicative of better wellbeing.

4.8.8. Descriptive analysis of the measures and sample

The utility, profile, dimension and total scores for each measure were assessed descriptively. This was done to understand the use of the measures for assessing the self-reported QoL of the sample. For the PBMs, a range of indicators including the mean (SD), median, and utility range were calculated. The proportion of values observed across the range of utilities were also calculated. By comparing utility values across measures, it is possible to understand how the same respondent's health is described and scored across instruments with different measurement perspectives. For the non-PBMs, the mean (SD), median and range of each dimension level or total score was calculated.

4.9. Results – Sample and measure descriptive statistics

4.9.1. Sample characteristics and survey completion process

Table 12 reports the overall survey completion process. In total, 907 respondents accessed the survey, 867 (95.6%) answered at least one question, and 794 (87.5%) fully completed the survey in more than the minimum time of four minutes. Each of the 24 survey blocks was completed between 28 and 41 times. The mean (median) time taken was 29.2 (21.8) minutes, with a range of 4.5 to 174.4. Observing the breakdown of time taken into intervals shows that 130 (16.4%) people took less than 15 minutes, and 73 (9.2%) took more than one hour to complete.

Table 12: Survey completion process

Category	N (%)
Full survey completer	794 (87.5)
Survey version completed 1	35 (4.4)
2	33 (4.2)
3	26 (3.3)
4	32 (4.0)
5	35 (4.4)
6	34 (4.3)
7	34 (4.3)
8	28 (3.5)
9	34 (4.3)
10	38 (4.8)
11	31 (3.9)
12	33 (4.2)
13	35 (4.4)
14	34 (4.3)
15	35 (4.4)
16	31 (3.9)
17	34 (4.3)
18	28 (3.5)
19	41 (5.2)
20	33 (4.2)
21	33 (4.2)
22	33 (4.2)
23	30 (3.8)
24	34 (4.3)
Time Taken (minutes)	
Mean (SD)	29.2 (12.2)
Range	4.5 to 174.4
Median	21.8
Time taken categories	
Less than 15 minutes	130 (16.4)
15 to 20	201 (25.3)
20.01 to 25	150 (18.9)
25.01 to 30	84 (10.6)
30.01 to 45	106 (14.6)
45.01 to 60	40 (5.0)
More than 60 minutes	73 (9.2)
Platform	
Windows	689 (86.8)
Macintosh	95 (12.0)

Table 13 reports the overall sample demographics. A large proportion of respondents (63.1%) reported experiencing at least one co-morbid health condition, with the most prevalent being pain (28.8%), depression (24.6%), anxiety (21.3%) and hypertension (21.0%). Tiredness was

reported by 27.4% of the sample. Overall, 55% of respondents report having between one and five co-morbid conditions or health concerns.

Table 13: Sample demographics

Category	N(%)	Category	N(%)
Overall	794	*Not hospitalised in last year*	612 (77.8)
Age		*Visits to GP in last year*	
18 – 29	128 (16.1)	0	62 (7.8)
30 – 44	202 (25.4)	1	86 (10.8)
45 – 59	222 (28.0)	2	115 (14.5)
60 – 74	220 (27.7)	3	108 (13.6)
75+	20 (2.5)	4	90 (11.3)
Gender		5	52 (6.6)
Male	380 (47.9)	6	72 (9.1)
Female	414 (52.1)	7 to 12	124 (15.6)
Country Of Birth		More than 12	85 (10.7)
Australia	623 (78.9)	Income	
Other	167 (21.1)	Less than 20,000	135 (17.0)
Health Conditions		20,001 to 30,000	147 (18.5)
Have any condition	500 (63.1)	30,001 to 40,000	70 (8.8)
Pain	228 (28.8)	40,001 to 50,000	67 (834)
Tiredness	217 (27.4)	50,001 to 60,000	67 (8.4)
Depression	195 (24.6)	60,001 to 70,000	57 (7.2)
Anxiety	169 (21.3)	70,001 to 80,000	42 (5.3)
High blood pressure	166 (21.0)	80,001 to 100,000	57 (7.2)
Insomnia	111 (14.0)	More than 100,000	77 (9.7)
Breathing problems	110 (13.9)	Prefer not to say	75 (9.5)
Diabetes	107 (13.5)	Marital status	
Arthritis	104 (13.1)	Married/de facto	465 (58.6)
Heart disease	40 (5.1)	Separated/divorced	97 (12.2)
Cancer	19 (2.4)	Single	210 (26.5)
Stroke	10 (1.3)	Widowed	22 (2.8)
Number Of Conditions		Education level	
0	292 (36.8)	Higher degree	71 (9.0)
1	93 (11.7)	Bachelor's degree	209 (26.3)
2	119 (15.0)	Trade certificate/diploma	247 (31.1)
3	107 (13.5)	Secondary	251 (31.6)
4	66 (8.3)	Primary	16 (2.0)
5	55 (6.9)	*Have children*	389 (49.0)
6 or more	60 (7.6)		

4.9.2. Descriptive analysis of measures

Table 14 reports descriptive statistics (the mean, median, standard deviation, frequency of utilities by severity category) for the value sets. Firstly, the mean EQ-5D-5L values were higher than the SF-6D. The ASCOT and ICECAP-A values are in a similar range to the EQ-5D-5L. This is

worth noting as although the overall utility values are similar, the diverse areas of QoL measured by the descriptive systems means that this is cannot be interpreted as evidence of convergence. Overall, 17.5% of respondents report that they were in the best health state described by the EQ-5D-5L with no problems on each dimension. This provides evidence of a ceiling effect, a response characteristic that is widespread in EQ-5D data [209]. The ICECAP-A and ASCOT report a similar number of responses below a utility value of 0.5 (11.2% and 10.4% respectively). No respondents reported being in the best ASCOT health state in comparison to 12.1% on the ICECAP-A.

Table 14: Descriptive statistics for each of the value sets estimated

Measure	Mean (SD)	Median	Value set range				
			N (%) Below 0	N (%) 0.001 to 0.500	N (%) 0.501 to 0.750	N (%) 0.751 to 0.999	N (%) at 1
EQ-5D-5L	0.699 (0.28)	0.754	18 (2.3)	143 (18.0)	221 (27.8)	273 (34.4)	139 (17.5)
SF-6D	0.509 (0.26)	0.519	31 (3.9)	337 (42.4)	263 (33.1)	150 (18.9)	13 (1.6)
ICECAP	0.774 (0.21)	0.849	4 (0.5)	85 (10.7)	186 (23.4)	423 (53.3)	96 (12.1)
ASCOT	0.769 (0.20)	0.814	3 (0.4)	79 (10.0)	202 (25.4)	510 (64.2)	0

Table 15 reports the descriptive statistics for the non preference-based HRQoL and wellbeing instruments. The SF-36 PF dimension had the highest overall score across the eight dimensions. The RP, RE and SF dimension scores were also generally high. The MH and VT dimension scores were lower indicating that the sample reported a higher prevalence of issues with mental health and energy. The GH dimension score, which is an indicator of the overall health of the sample, was also in the moderate range w. The mean PROMIS-29 dimension scores were between 47.3 and 53.3, and were therefore broadly in line with the average for the United States population. WEMWBS scores demonstrated that the sample reported generally good wellbeing.

Table 15: SF-36 and PROMIS dimension, WEMWBS and ONS-4 scores

Measure and dimension		Mean (SD)	Median	Range
SF-36[a]				
	PF	74.0 (26.1)	85.0	0 to 100
	RP	70.4 (26.0)	75.0	0 to 100
	RE	72.5 (26.5)	83.3	0 to 100
	SF	72.7 (26.6)	75.0	0 to 100
	PA	67.4 (24.7)	66.7	0 to 100
	MH	64.9 (21.9)	65.0	5 to 100
	VT	49.5 (22.2)	50.0	0 to 100
	GH	54.5 (21.7)	55.0	0 to 100
PROMIS-29[b]				
	PF	47.3 (8.5)	48.0	22.9 to 56.9
	A	50.6 (14.6)	53.7	48.0 to 81.6
	D	53.3 (10.1)	53.9	41.0 to 79.4
	FA	53.0 (10.2)	51.0	33.7 to 75.8
	SL	50.5 (10.0)	48.4	32.0 to 73.3
	SF	51.5 (8.5)	51.6	33.6 to 64.1
	PA	52.6	53.9	41.6 to 75.6
WEMWBS				
	Total Score	46.2 (10.9)	46	14 to 70

[a] PF: Physical Functioning; RP: Role Physical; RE: Role Emotional; SF: Social Functioning; PA: Pain; MH: Mental Health; VT: Vitality; GH: General Health. All dimensions have possible 0 to 100 score

[b] PF: Physical Function (possible range 22.9-56.9); A: Anxiety (range 40.3-81.6); D: Depression (range 41.0-79.4); FA: Fatigue (range 33.7-75.8); SL: Sleep (32.0-73.3); SF: Social Function (range 27.5-64.2); PA: Pain inference (range 41.6-75.6)

[c] WEMWBS possible score range 14 to 70

4.10. Extension 1 – Data analysis

EFA [210] was used to examine the dimensionality of the pooled item responses from the QoL outcome measures included. A number of EFA models with a variety of specifications were tested to understand the most appropriate dimension structure. EFA tests the dimensionality of groups of items without imposing a pre-specified factor structure, and is used to understand the latent structure of tests.

Given the aim of this study was to investigate the relationship between diverse concepts of QoL, and the broadening of QoL measurement, EFA was preferred to the use of confirmatory factor analysis (CFA). This is because CFA requires a predefined factor structure to be specified for the data, with model fit statistics estimated for the data fitting the proposed model. As this work assessed different dimensionalities from a broad item pool with no predefined dimension structure, EFA was used.

4.10.1. Estimation procedures

To estimate the factor structure, a maximum likelihood approach was used. Maximum likelihood estimations assume multivariate normal observations (but can be used for both normal and non-normal data) and maximise the determinant of the correlation matrix. A maximum likelihood approach was used as it is able to estimate models including large numbers of items and more than three dimensions.. A further advantage is that maximum likelihood approaches have been shown to accurately estimate the necessary factor correlation matrix from which the factor structure is drawn [196, 197]. The maximum likelihood approach used to extract factors was the Metropolis–Hastings Robbins-Munro (MH-RM) algorithm. This algorithm was proposed for use in EFA by Cai [211,212]. It uses a fixed simulation size throughout a series of iterations to converge to a local maximum and subsequently estimate the factor structure for the item pool.

4.10.2. Model specifications

The item pool tested in the EFA analyses included the majority of the items from the measures, with a number of exclusions. Overall, 91 items were included in the modelling. These were five items from the EQ-5D-5L, nine from the ASCOT, five from the ICECAP-A, 30 from the SF-36, 28 from the PROMIS-29, and 14 from the WEMWBS. The five SF-36 GH dimension questions were excluded as the aim of this study was to test the relationship of items across different areas of specific health and wider QoL, rather than including general health indicators. The ONS-4 and the PROMIS pain intensity items were also excluded as these are continuous VAS type scales so therefore differ to the categorical response questions used elsewhere.

A range of factor structures were tested, with models including an unrestricted number of dimensions estimated for all specifications. Models including between 6 and 10 dimensions were also tested.

The stability of the models produced for the full sample was also tested for each model specification. To allow for this comparison, five subsamples of approximately half the sample in the data stratified by age and gender were randomly selected using a random number generator approach. Each model specification was then tested on each of the subsamples, and the stability of the model structures in comparison to the model produced for the overall sample was examined.

Table 16 reports the EFA models tested. This was based on an iterative approach to selecting the preferred model, where firstly both oblique and orthogonal rotation were tested, and a decision was made about which to use. This decision was based on the interpretability of the

dimension structures. Each of the structures was assessed for consistency and comprehensibility for use in unidimensional IRT analyses.

Within the rotated factor structures, items were included in a factor if the factor loading was at a level of 0.3 or above. Past work has used higher minimum correlations such as 0.32 or 0.4 [213]. However, a lower level was used here to generate inclusive dimensions. Items with factor loadings above 0.30 on more than one dimension [214], but within 0.2 across the correlations [215] were also examined as potential cross loaders, but were retained in factor models.

The stability of the models produced for the full sample was also tested for each model specification. To allow for this comparison, five subsamples of approximately half the sample in the data stratified by age and gender were randomly selected using a random number generator approach. Each model specification was then tested on each of the subsamples, and the stability of the model structures in comparison to the model produced for the overall sample was examined.

Table 16: EFA models tested

Model	Estimation	Factor number extracted	Rotation method	Rotation criteria	Justification
Model 1	MH-RM[a]	Unrestricted (results in 10)	Oblique	CF-Quartimax[b]	Compare rotation method, assess correlations and choose which to proceed with
Model 2	MH-RM	Unrestricted (results in 10)	Oblique	CF-Varimax[c]	
Model 3	MH-RM	Unrestricted (results in 11)	Orthogonal	CF-Quartimax	
Model 4	MH-RM	Unrestricted (results in 11)	Orthogonal	CF-Varimax	
Model 5	MH-RM	10	Oblique	CF-Quartimax	Assess models across both rotation criteria, and selected the dimension structure for the next stages
Model 6	MH-RM	10	Oblique	CF-Varimax	
Model 7	MH-RM	9	Oblique	CF-Quartimax	
Model 8	MH-RM	9	Oblique	CF-Varimax	
Model 9	MH-RM	8	Oblique	CF-Quartimax	
Model 10	MH-RM	8	Oblique	CF-Varimax	
Model 11	MH-RM	7	Oblique	CF-Quartimax	
Model 12	MH-RM	7	Oblique	CF-Varimax	
Model 13	MH-RM	6	Oblique	CF-Quartimax	
Model 14	MH-RM	6	Oblique	CF-Varimax	

[a] Metropolis–Hastings Robbins-Munro algorithm; [b] Crawford Ferguson Quartimax; [c] Crawford Ferguson Varimax

4.10.3. Rotation method and criteria

Both orthogonal and oblique factor rotation methods were tested for use in the identification of the dimension structure. Oblique rotation assumes correlations between the factors, where orthogonal rotation assumes independence. Oblique rotation models also generally have more interpretable factors, with a simpler structure than that obtained with an orthogonal rotation. There is a potentially stronger theoretical position for using oblique rotations which is based on the relationship between QoL concepts as experienced within an individual, or across a group setting. As there is likely to be a relationship between QoL concepts measured, this supports the assertion that dimensions of QoL are correlated.

The factor rotation criteria tested included the Crawford Ferguson (CF) Quartimax and CF-Varimax methods. CF criteria have been described as the most comprehensive group of rotations in the identification of dimensionality [216]. CF-Varimax aims to maximise the variance of the squared loadings in each factor by distributing the variance across the factor structure. This means that each factor often includes loading at either a large or small level, therefore facilitating the identification of each variable as belonging to a different factor. CF-Quartimax aims to minimise the complexity of the variable relationships, and therefore functions well when the data has distinct clusters and few items cross loading between factors. CF-Varimax and Quartimax methods with oblique rotation have been compared empirically and were found to produce similar point estimates for rotated factor loadings and factor correlations but varying standard error estimates [216].

4.10.4. Generating a dimension structure for further testing

The dimensionality established for Extension 1 was used as the basis for the unidimensional item groups tested for Extension 2. The requirements of the dimensions taken forward for testing were that they included a PBM item (that could be conceptualised as the preference-based Layer 1). Given the widespread use of the EQ-5D-5L, and the variety of dimensions covered, how those five items aligned with the wider dimension structure, and their amenability to testing for the layered approach as part of Extension 2, was examined.

4.11. Results - Dimensionality assessment

Table 16 displays the EFA model criteria combinations tested on the overall item pool. **Model 1** to **Model 4** are unrestricted in terms of the number of dimensions estimated, and the other models are specified to have a certain number of dimensions. **Table 17** and **Table 18** report the factor structure (also with factor loadings and standard errors) resulting from the use of oblique

rotation, with both quartimax (**Model 1**) and varimax (**Model 2**) methods. Appendix 8 reports the factor structure (with factor loadings greater than 0.3) from the unrestricted quartimax (**Model 3**) and varimax (**Model 4**) orthogonal models. The tables use shorthand codes for the items, and these codes, and the associated full item descriptions, are provided in Appendix 9.

The dimensionality of the orthogonal models resulted in a single factor including the majority of the item pool, with cross loading between dimensions evident. In contrast, the oblique models did not produce a single factor combining many items, and had a clear structure resulting in conceptually interpretable dimensions (with minor differences between models). The unrestricted models resulted in an 11 dimension structure. Testing models with less dimensions resulted in combinations of the 11, with increased cross loading. This increased the difficulty of defining the factor structures.

These results suggested that the 11-dimension structure using oblique varimax and quartimax rotation explained the data in the clearest and most interpretable way. Analysis of the stability of the models across the five randomly selected subsamples found that the 11 dimension model structures were generally stable, and interpretable in the same structure as the overall model. Therefore, this structure was used to examine how these instruments facilitate the measurement of wider outcomes, and a potential layered approach to measurement. Further detail about the characteristics of each dimension is provided below (See also **Table 17** and **Table 18**, and Appendix 9 which describes the full text for the item coding used in the tables):

- *Dimension 1*: Defined as a mental health factor, and included all 14 items measuring mental health issues from the EQ-5D-5L, SF-36 and PROMIS-29.
- *Dimension 2*: A composite of 13 wider QoL concepts that included seven SCRQoL items from the ASCOT, all five ICECAP-A items assessing capabilities, and one WEMWBS item asking about frequency of feeling loved (which is in line with the capability approach).
- *Dimension 3:* Defined as a physical functioning factor and included 17 items. Both models included EQ-5D-5L MO, with the quartimax model also including EQ-5D-5L SC and UA. Consistent across both models was the inclusion of the ten PF dimension items from the SF-36, and the four PF items from the PROMIS-29.
- *Dimension 4:* Defined as a sleep dimension, as it included the four PROMIS-29 items about sleep and sleep quality. This is a concept not assessed by the PBMs included in the study.

- *Dimension 5:* Consistently included the two ASCOT items investigating dignity. The quartimax model also included two WEMWBS items asking about the frequency of feeling relaxed, and of being able to make up your mind about things. The varimax model also included two WEMWBS items asking about the frequency of dealing with problems well and thinking clearly.
- *Dimension 6*: Defined as a tiredness/energy dimension. It consistently included the two negative vitality items from the SF-36v2, and the four PROMIS-29 tiredness items. The quartimax model also included an energy item from the WEMWBS. Energy is a concept included in the SF-6D, but not the EQ-5D-5L.
- *Dimension 7:* A role and social functioning dimension that included only SF-36 items. Role functioning is a dimension included in the SF-6D. This model demonstrates that it differs to the activity based items included in other instruments.
- *Dimension 8:* Included four positively worded energy and mental health items from the SF-36, and one energy item from the WEMWBS (which cross loads with dimension 6).
- *Dimension 9:* Included nine WEMWBS items covering positive wellbeing concepts based around asking about feelings. The varimax model also included the ICECAP item assessing love friendship and support which cross loads with dimension 2.
- *Dimension 10:* Defined as a pain dimension. It consistently included EQ-5D-5L PD, the two SF-36v2 and four PROMIS-29 pain items. The varimax model also included the PROMIS-29 item assessing respondent ability to do chores. Pain appears in both the generic HRQoL PBMs included in this study.
- *Dimension 11:* Consistently included EQ-5D-5L SC, and two ASCOT questions about appearance, and access to food and drink. The varimax model also included the SF-36v2 item about bathing and dressing, and two PROMIS walking items. This indicated overlap in terms of coverage of self-care, and appearance, but broader concepts were also included.
- Non-loading items (i.e. loading below 0.3 across all dimensions) can also be conceptualised as measuring social and wider activities. This is because the four PROMIS activities items were consistently non-loading. The varimax model also includes EQ-5D-5L usual activities and an SF-36v2 social activities item as non-loaders.

4.12. Broadening the standard framework

The results of the factor analysis demonstrated how the standard HRQoL framework fits with, and could be extended by, the addition of other constructs of QoL. The combination of QoL outcomes included in this study could generate a wider measure if existing measures were used as the basis for its development.

Results worth noting in terms of expanding the framework include the identification of a clear dimension structure linked to the EQ-5D-5L, where the 'functioning' dimensions (MO, SC and UA) clustered together in some models, with the 'symptoms' dimensions (PD and AD) operating as separate factors. Beyond the EQ-5D framework, the results suggested a composite dimension assessing SCRQoL and capabilities. There was also evidence for dimensions that clustered based on broader constructs measured (for example sleep), and some evidence for clustering based on the direction in which the construct is assessed (i.e. positively or negatively worded).

Table 17: Dimension structure - EFA Oblique quartimax rotation (Model 1)

Dimension 1		Dimension 2		Dimension 3	
Item	**FL (SE)[a]**	**Item**	**FL (SE)**	**Item**	**FL (SE)**
EQ AD	0.60 (0.03)	ASCOT 1	0.57 (0.04)	EQ MO	0.67 (0.02)
SF-36 Q24	0.46 (0.05)	ASCOT 2	0.32 (0.04)	EQ SC	0.45 (0.02)
SF-36 Q25	0.49 (0.04)	ASCOT 3	0.38 (0.05)	EQ UA	0.34 (0.02)
SF-36 Q26	0.31 (0.05)	ASCOT 4	0.44 (0.06)	SF-36 3	0.69 (0.06)
SF-36 Q28	0.47 (0.04)	ASCOT 5	0.70 (0.05)	SF-36 4	0.62 (0.06)
SF-36 Q30	0.31 (0.05)	ASCOT 6	0.66 (0.05)	SF-36 5	0.67 (0.06)
PROMIS Q5	0.53 (0.05)	ASCOT 7	0.35 (0.07)	SF-36 6	0.87 (0.09)
PROMIS Q6	0.59 (0.04)	ICECAP 1	0.56 (0.05)	SF-36 7	0.88 (0.05)
PROMIS Q7	0.57 (0.04)	ICECAP 2	0.64 (0.07)	SF-36 8	0.70 (0.06)
PROMIS Q8	0.56 (0.04)	ICECAP 3	0.48 (0.07)	SF-36 9	0.82 (N/R)
PROMIS Q9	0.53 (0.04)	ICECAP 4	0.63 (0.09)	SF-36 10	0.80 (0.07)
PROMIS Q10	0.49 (0.03)	ICECAP 5	0.62 (0.05)	SF-36 11	0.73 (0.06)
PROMIS Q11	0.62 (0.03)	WEMWBS 12	0.33 (0.06)	SF-36 12	0.51 (0.08)
PROMIS Q12	0.56 (0.03)			PROMIS 1	0.45 (0.05)
				PROMIS 2	0.69 (0.06)
				PROMIS 3	0.65 (0.05)
				PROMIS 4	0.52 (0.06)
Dimension 4		**Dimension 5**		**Dimension 6**	
Item	**FL (SE)**	**Item**	**FL (SE)**	**Item**	**FL (SE)**
PROMIS 17	0.68 (0.04)	ASCOT 8	0.46 (N/R)	SF-36 29	0.62 (0.05)
PROMIS 18	0.63 (0.03)	ASCOT 9	0.35 (0.04)	SF-36 31	0.73 (0.05)
PROMIS 19	0.70 (0.04)	WEMWBS 3	0.34 (0.04)	PROMIS 13	0.89 (0.04)
PROMIS 20	0.58 (0.03)	WEMWBS 11	0.40 (0.04)	PROMIS 14	0.91 (N/R)
				PROMIS 15	0.79 (0.03)
				PROMIS 16	0.90 (0.04)
				WEMWBS 5	0.37 (0.05)
Dimension 7		**Dimension 8**		**Dimension 9**	
Item	**FL (SE)**	**Item**	**FL (SE)**	**Item**	**FL (SE)**
SF-36 13	0.61 (0.08)	SF-36 23	0.49 (0.04)	WEMWBS 1	0.41 (0.02)
SF-36 14	0.78 (0.13)	SF-36 26	0.36 (0.05)	WEMWBS 2	0.36 (0.04)
SF-36 15	0.69 (0.05)	SF-36 27	0.49 (0.09)	WEMWBS 4	0.42 (0.05)
SF-36 16	0.68 (0.06)	SF-36 30	0.39 (0.04)	WEMWBS 8	0.39 (0.03)
SF-36 17	0.73 (0.05)	WEMWBS 5	0.30 (0.05)	WEMWBS 9	0.51 (0.04)
SF-36 18	0.68 (0.06)			WEMWBS 10	0.39 (0.03)
SF-36 19	0.57 (0.07)			WEMWBS 12	0.48 (0.04)
SF-36 20	0.30 (0.06)			WEMWBS 13	0.41 (0.05)
SF-36 32	0.37 (0.05)			WEMWBS 14	0.43 (0.03)
Dimension 10		**Dimension 11**		**Non-loaders[a]**	
Item	**FL (SE)**	**Item**	**FL (SE)**		
EQ PD	0.68 (0.02)	EQ SC	0.37 (0.02)	PROMIS 21	N/A[c]
SF-36 21	0.75 (0.05)	ASCOT 2	0.38 (0.03)	PROMIS 22	N/A
SF-36 22	0.66 (0.04)	ASCOT 3	0.35 (0.03)	PROMIS 23	N/A
PROMIS 25	0.80 (0.07)			PROMIS 24	N/A
PROMIS 26	0.78 (0.03)			WEMWBS 6	N/A
PROMIS 27	0.64 (0.04)			WEMWBS 7	N/A
PROMIS 28	0.73 (0.03)				

[a] Factor loading (standard error) [b] Items not loading on any dimension; Note: fuller item descriptions in Appendix 9; [c] not applicable

Table 18: Dimension structure - EFA Oblique Varimax rotation (Model 2)

Dimension 1		Dimension 2		Dimension 3	
Item	**FL (SE)[a]**	**Item**	**FL (SE)**	**Item**	**FL (SE)**
EQ AD	0.56 (0.02)	ASCOT 1	0.55 (0.03)	EQ MO	0.49 (0.01)
SF-36 24	0.42 (0.05)	ASCOT 3	0.34 (0.04)	SF-36 3	0.64 (0.06)
SF-36 25	0.45 (0.04)	ASCOT 4	0.42 (0.05)	SF-36 4	0.53 (0.06)
SF-36 28	0.44 (0.03)	ASCOT 5	0.66 (0.05)	SF-36 5	0.56 (0.06)
PROMIS 5	0.49 (0.04)	ASCOT 6	0.63 (0.05)	SF-36 6	0.77 (0.08)
PROMIS 6	0.54 (0.04)	ASCOT 7	0.31 (0.07)	SF-36 7	0.74 (0.05)
PROMIS 7	0.53 (0.04)	ICECAP 1	0.53 (0.05)	SF-36 8	0.62 (0.05)
PROMIS 8	0.53 (0.04)	ICECAP 2	0.58 (0.07)	SF-36 9	0.67 (0.04)
PROMIS 9	0.49 (0.04)	ICECAP 3	0.45 (0.07)	SF-36 10	0.62 (0.08)
PROMIS 10	0.45 (0.03)	ICECAP 4	0.59 (0.09)	SF-36 11	0.56 (0.06)
PROMIS 11	0.58 (0.03)	ICECAP 5	0.58 (0.05)	SF-36 12	0.35 (0.07)
PROMIS 12	0.51 (0.03)			PROMIS 1	0.30 (0.05)
				PROMIS 2	0.50 (0.05)
				PROMIS 3	0.42 (0.06)
				PROMIS 4	0.32 (0.06)
Dimension 4		**Dimension 5**		**Dimension 6**	
Item	**FL (SE)**	**Item**	**FL (SE)**	**Item**	**FL (SE)**
PROMIS 17	0.70 (0.04)	ASCOT 8	0.49 (N/R)	SF-36 29	0.58 (0.05)
PROMIS 18	0.63 (0.03)	ASCOT 9	0.37 (0.11)	SF-36 31	0.68 (0.05)
PROMIS 19	0.73 (0.05)	WEMWBS 3	0.38 (0.04)	PROMIS 13	0.84 (0.04)
PROMIS 20	0.61 (0.06)	WEMWBS 6	0.32 (0.04)	PROMIS 14	0.85 (N/R)
		WEMWBS 7	0.31 (0.05)	PROMIS 15	0.74 (0.02)
		WEMWBS 11	0.45 (0.04)	PROMIS 16	0.85 (0.04)
Dimension 7		**Dimension 8**		**Dimension 9**	
Item	**FL (SE)**	**Item**	**FL (SE)**	**Item**	**FL (SE)**
SF-36 13	0.59 (0.07)	SF-36 23	0.54 (0.04)	ICECAP 2	0.31 (0.04)
SF-36 14	0.75 (0.12)	SF-36 26	0.41 (0.05)	WEMWBS 1	0.41 (0.02)
SF-36 15	0.67 (0.05)	SF-36 27	0.55 (0.09)	WEMWBS 2	0.35 (0.04)
SF-36 16	0.66 (0.06)	SF-36 30	0.44 (0.04)	WEMWBS 4	0.42 (0.06)
SF-36 17	0.70 (0.05)	WEMWBS 5	0.34 (0.05)	WEMWBS 8	0.39 (0.03)
SF-36 18	0.66 (0.05)			WEMWBS 9	0.53 (0.04)
SF-36 19	0.55 (0.06)			WEMWBS 10	0.40 (0.03)
SF-36 32	0.36 (0.05)			WEMWBS 12	0.51 (0.04)
				WEMWBS 13	0.41 (0.05)
				WEMWBS 14	0.44 (0.03)
Dimension 10		**Dimension 11**		**Non-loaders[b]**	
Item	**FL (SE)**	**Item**	**FL (SE)**		
EQ PD	0.66 (0.02)	EQ MO	0.30 (0.01)	EQ UA	N/A[c]
SF-36 21	0.72 (0.04)	EQ SC	0.49 (0.02)	PROMIS 21	N/A
SF-36 22	0.63 (0.04)	ASCOT 2	0.42 (0.03)	PROMIS 22	N/A
PROMIS 1	0.31 (0.05)	ASCOT 3	0.38 (0.04)	PROMIS 23	N/A
PROMIS 25	0.76 (0.06)	SF-36 12	0.32 (0.06)	PROMIS 24	N/A
PROMIS 26	0.74 (0.03)	PROMIS 3	0.43 (0.14)	SF-36 20	N/A
PROMIS 27	0.60 (0.04)	PROMIS 4	0.41 (0.06)		
PROMIS 28	0.69 (0.03)				

[a] Factor loading (standard error) [b] Items not loading on any dimension; Note: fuller item descriptions in Appendix 9; [c] not applicable

4.13. Establishing a dimensionality for Extension 2 analyses

The results of the EFA analysis provided a basis for the testing a layered approach to measurement. This could be approached using EQ-5D-5L items as the preference-based layer to enable a dimension structure for the IRT analysis to be conceptualised. Dimensions 1, 3 and 10 described in the results above were defined as measuring mental health, physical functioning and pain, and all include EQ-5D-5L items that could be conceptualised as the preference-based layer. These dimensions were tested using IRT methods. The non-loading activities items also included an EQ-5D-5L item, and were therefore also tested as a separate dimension. The dimension that included EQ-5D-5L SC was not tested given the inconclusive nature of that factor which only included a small number of items, limit the conclusions that could be drawn from the results.

4.14. Extension 2 – Investigating a layered approach to measurement

In the analysis for Extension 2, IRT indicators were used to examine the relationship between, and the performance of, the items included. Given that the aim of this work was to examine the item pools, and not develop a new measure, all items were retained in the models. The implications of the IRT analysis for the use of the items in a layered approach to measurement, was examined. IRT was conducted using IRTPRO [217], a specialist software for this purpose.

4.15. Description of general IRT approach

4.15.1. Which IRT model was used?

The two parameter graded response IRT model [178] was used in this study. This model can be applied when item severity level responses have a logical ordering of severity and are polytomous. This is the case for the items included in this study. The graded response model has also been previously used for assessing HRQoL outcome measures (as described in Section 4.4.5). The two parameter model has the flexibility to include items with different response formats (for example both severity and frequency) on the same theta scale [190]. This is important in this study, as the items are taken from different questionnaires, which are likely to have different discrimination profiles given that the wording and response levels used differ.

In this study the theta scale describes different unidimensional aspects of QoL with high and low severity levels. Alongside the information provided by the two parameters, the models were also used to evaluate other item indicators including local dependence, DIF, and model fit.

4.15.2. How is IRT used in this study?

In this study, the IRT conducted builds on previous use of the graded response model, [218] and applies it in a novel way. Models are applied to combined item pools from the instruments included to assess item performance, and the information provided, and demonstrate how a layered approach to measuring outcomes could be operationalised. There is limited work using IRT methods to assess existing measures of QoL in comparison to each other, and this is a novelty of the work conducted here. This application of IRT differs from the original purpose of IRT to develop and refine precise measurement metrics, and is more exploratory.

It is often the case that IRT models are iteratively used to refine the items within each dimension. The exploratory nature of this work means that the overall graded response model produced for each dimension will be assessed to understand the relationship between items, violations of the assumptions of IRT, and individual item performance. IRT results will not be iteratively used to refine the items included in each dimension, but will be used to indicate where refinements could be made. This means that the results inform the future development of each extension tested, rather than producing a final product.

4.16. Data analysis process

The IRT process was informed by a series of steps outlined in existing guidance papers [190]. This included guidance by Toland [219] who described a multistep process for the application of IRT to patient-reported outcome data. The approach was adapted for use in this study. A series of steps was undertaken for each dimension structure tested.

The first step involved preliminary data inspection followed by item calibration (to allow the rest of the analyses to be conducted). Common IRT assumptions including local independence, model and item fit, and DIF were assessed. This was followed by an examination of the item thresholds, information curves, and slope parameters, and an assessment of the overall information provided by the items in each dimension. Finally, calibrated theta scores for the overall dimension (which could be used to facilitate CAT approaches) were examined. The results were used to assess the feasibility of developing a layered approach to measuring QoL outcomes for each of the dimensions tested. Each of these steps is now described.

4.16.1. Preliminary descriptive analysis – Data inspection

Prior to conducting IRT, it is recommended that descriptive analysis of responses across each of the item severity levels is conducted. This was done for each item set included in each dimension to test the distribution of data across response categories. Distribution is important as data is required at each response level to estimate IRT models with confidence.

4.16.2. Test common IRT assumptions – Local independence

The results of the item set calibration were used to test the key IRT assumptions. These assumptions are also indicators of item performance. The first of assumption is local independence between items within a dimension. This specifies that item responses should be independent of each other after controlling for the underlying construct. To test local independence, Standardised Chi Square values for each item pair were calculated. Values of greater than ten were considered potentially large, and possibly indicative of a dependency issue. Overlap in the content was observed, and the wording was assessed for qualitative redundancy.

4.16.3. Test functional form and model-data fit

Functional form implies that all threshold parameters are ordered, and there is a common slope within each item, although not necessarily across items. Model-data fit at the item level was tested, and model fit statistics were used to compare the relative fit of the overall model.

Item level fit:

To assess the fit at the item level, the graded response adaptation of S-χ2 diagnostic fit statistic [220] was used. This statistic assesses the degree of similarity between the predicted and observed response frequencies for each item. A statistically significant item value ($p<0.01$) indicates that the item does not fit the model.

Model level fit:

A range of overall model level fit statistics, which compare the relative fit of the model to the data, were tested. These included the AIC [221] and the BIC [222]. The AIC and BIC are based on in-sample fit and estimates the likelihood of a model for estimating the specified values. The BIC is fit measure that is a function of both the number of parameters and the number of observations, and provides a better test of complex models. Lower values are indicative of better fit, but significance cannot be calculated.

The M_2 goodness-of-fit statistic measures the fit of the model to the data assuming perfect model fit. This produces a significance estimate and non-significance is preferred. The M_2 test is sensitive to misfit to the underlying dimension. The Root Mean Square Error of Approximation (RMSEA), which assesses fit including an adjustment for sample size, was also estimated. The RMSEA ranges from 0 to 1, with smaller values preferred.

4.16.4. Testing for DIF

DIF is the assessment of differences in item performance between subgroups (for example demographic groups) at different points of the theta scale. It exists when the probabilities of item endorsement are different across subgroups with the same level of the underlying trait. DIF is an issue as it means that item responses may be reflecting characteristics of the subgroup rather than measuring the underlying trait.

In this study, an omnibus Wald Chi Square test was used. In the initial test, all of the items within a dimension were used to estimate the mean level of the underlying trait for each subgroup. The item parameters were then estimated separately for each subgroup to detect DIF. However, for some subgroups, there were not responses for every level of all items. If this occurred these items were removed from the analysis. Significant differences in parameter estimates across the groups for each item were identified using $p<0.01$ as an indicator of significance. This was done for all items included in each dimension (where each subgroup appeared in each item response category).

The subgroups tested were gender and whether or not the respondent reported having a long-term condition. However, the presence of DIF between these groups could be argued to have different levels of interpretation. For example, DIF by gender could be an indicator of measuring between gender differences rather than the underlying trait. However, DIF by long-term condition status may be an indicator of sensitivity to a particular condition which may be a psychometric advantage. Both were included in the analysis, with the results interpreted accordingly.

4.16.5. Item level thresholds

Item characteristic curves were assessed to ensure that the response levels were operating as expected for each item. The threshold parameters were also compared across the items to understand which items provided the widest coverage of the underlying theta scale, and

whether certain items provided more information (and therefore more sensitive measurement abilities) at different points of theta.

4.16.6. Item level information and the total information provided

The slope parameter magnitudes were compared across items. The individual item information curves were assessed to understand the characteristics of the information provided by each item (see Section 4.6.1 for information about how these are calculated). These curves help understand where each item is contributing information. They are a function of the estimated slope and threshold parameters, where the information provided for each item is highest at the points of theta where the threshold parameter values occur.

The total information function (TIF) curves presenting the cumulative information for all items across theta were also evaluated to investigate of the information provided at different severity points of the theta scale. This is calculated as the aggregate of the individual item function curves, so is directly linked to the profile of the information provided by each item. Alongside this, the standard error of the estimates (calculated as 1/√information) is calculated to provide an indicator of the amount of error in scores across theta. This is useful to understand the precision of estimated theta scores as a function of the information provided at different levels of the latent trait. A scale that is sensitive to the full range of the underlying trait would have a similar level of information and a low standard error across the latent trait. In this analysis, the TIF and standard error curves are plotted, and the values of each at certain points of theta are presented as an indicator of overall scale sensitivity.

4.16.7. Estimated calibrated dimension score curve

The expected score curve (or test characteristic curve) shows the expected theta score that equates to a summed score from an item set (and could be used in the facilitation of a CAT based approach). It links item total scores to calibrated theta scores. Raw item scores were summed to generate a dimension score, and this was mapped onto the theta scores. This allowed for an assessment of the raw dimension score levels required at certain points of the calibrated scale. It also demonstrates the characteristics of the item information provided (in terms of increases in the raw score required to have the same change on the theta scale at different severity points). These scores could be assessed for individuals and compared to calibrated scores across other dimensions.

4.17. Results overview

The sections below report the IRT results for the four dimensions brought forward from Extension 1. Each dimension includes at least one EQ-5D-5L dimension that could be conceptualised as a preference-based dimension in a multilayered instrument. The dimensions have been defined as physical functioning, mental health, pain, and activities. As the IRT analyses are exploratory, one model is calibrated for each dimension, with the assumptions of the IRT model, and item and overall dimension performance indicators assessed.

4.18. Results – Physical Functioning dimension

4.18.1. Justification of dimensionality

The items included in the physical functioning dimension were based on the dimensionality highlighted by the EFA using the oblique quartimax rotation model. This dimensionality was preferred as it includes the largest number of items related to mobility and physical functioning. This was seen as advantageous given the aims of the analysis were to explore the relationship between items measuring similar concepts to investigate their use in the assessment of wider concepts as part of a layered approach to measuring health and QoL outcomes.

4.18.2. Initial data inspection

Table 19 reports the frequency of respondents in each category. The SF-36 and PROMIS-29 item scores were reversed so that a low score indicates a low level of problems in line with the EQ-5D-5L, which was used as the reference frame given the wider aims of the analysis. Column L1 indicates the least severe level of problems for all items. There were limited responses at level 5 (unable to/extreme problems) of the EQ-5D-5L dimensions. Small numbers have in past work been suggested as sufficient for analysis [219]. The responses were sufficiently distributed across the other severity levels.

4.18.3. Assessing local independence

Standardised Chi Square values for each pair of items were assessed. **Table 20** describes the item pair correlations above the indicative cut off of 10, and Appendix 10 includes all of the estimated Chi Square values for each item pair. Overall, 13 (9.5%) of pairs had a value above 10. PROMIS item 1 (ability to do chores) and item 2 (up and down stairs at normal pace) exhibit dependency with four and three SF-36 items respectively. Although there is some evidence of local dependence, examining the item pairs suggests some qualitative differences which may support inclusion of both items. An example of an item pair with observable local dependency

are the items 'walking more than a kilometre', and 'walking several hundred metres', both of which are taken from the SF-36.

Table 19: Initial data inspection - Physical functioning dimension

Item	L1[a]	L2[b]	L3[c]	L4[d]	L5[e]
EQ-5D-5L MO	485 (61.1)	186 (23.4)	93 (11.7)	25 (3.1)	2 (0.5)
EQ-5D-5L SC	667 (84.0)	72 (9.1)	44 (5.5)	6 (0.8)	5 (0.6)
EQ-5D-5L UA	439 (55.3)	224 (28.2)	102 (12.8)	24 (3.0)	5 (0.6)
SF-36 Q3	216 (27.2)	343 (43.2)	235 (29.6)	N/A	N/A
SF-36 Q4	450 (56.7)	261 (32.9)	83 (10.5)	N/A	N/A
SF-36 Q5	506 (63.7)	220 (27.7)	68 (8.6)	N/A	N/A
SF-36 Q6	378 (47.6)	291 (36.6)	125 (15.7)	N/A	N/A
SF-36 Q7	545 (68.6)	189 (23.8)	60 (7.6)	N/A	N/A
SF-36 Q8	374 (47.1)	308 (38.8)	112 (14.1)	N/A	N/A
SF-36 Q9	457 (57.6)	212 (26.7)	125 (15.7)	N/A	N/A
SF-36 Q10	547 (68.9)	159 (20.0)	88 (11.1)	N/A	N/A
SF-36 Q11	627 (79.0)	129 (16.2)	38 (4.8)	N/A	N/A
SF-36 Q12	671 (84.5)	97 (12.2)	26 (3.3)	N/A	N/A
PROMIS Q1	361 (45.5)	259 (32.6)	100 (12.6)	44 (5.5)	30 (3.8)
PROMIS Q2	408 (51.4)	204 (25.7)	96 (12.1)	52 (6.5)	34 (4.3)
PROMIS Q3	513 (64.6)	140 (17.6)	85 (10.7)	29 (3.7)	27 (3.4)
PROMIS Q4	553 (69.6)	132 (16.6)	69 (8.7)	25 (3.1)	15 (1.9)

[a] severity level 1 (least severe); [b] severity level 2; [c] severity level 3; [d] severity level 4; [e] severity level 5 (most severe)

Table 20: Item pair descriptors with a dependency > 10 – Physical functioning dimension

Item one	Item two	X^{2a}
SF-36 4: moderate activity limitations	PROMIS 2: Up and down stairs at normal pace	17.0
SF-36 7: Climbing one flight of stairs	PROMIS 2: Up and down stairs at normal pace	17.0
SF-36 6: Climbing several flights of stairs	PROMIS 1: Do chores	14.9
SF-36 4: Moderate activity limitations	PROMIS 1: Do chores	14.5
SF-36 5: Lifting or carrying groceries	PROMIS 2: Up and down stairs at normal pace	14.1
EQ Self-Care	SF-36 12: Bathing or dressing yourself	14.1
SF-36 5: Lifting or carrying groceries	PROMIS 1: Do chores	12.9
SF-36 4: Moderate activity limitations	SF-36 5: Lifting or carrying groceries	12.7
SF-36 7: Climbing one flight of stairs	PROMIS 1: Do chores	12.3
SF-36 6: Climbing several flights of stairs	PROMIS 4: Run errands and shop	11.8
SF-36 3: Vigorous activity limitations	SF-36 4: Moderate activity limitations	11.7
PROMIS 3: Walk of at least 15 minutes	PROMIS 4: Run errands and shop	11.7
SF-36 9: Walking more than a kilometre	SF-36 10: Walking several hundred metres	10.4

[a] Standardised Chi Square

4.18.4. Assessing model-data fit – Item level

The item level fit values are reported in the two right hand columns of **Table 21**. SF-36 Q12 (assessing ability to bathe and dress) was significant ($p < 0.001$) which indicated poor fit to the underlying model. This may be expected given the concept assessed which differs to the other items which are more focused on different areas of physical functioning.

4.18.5. Assessing model-data fit – Model level

The model level fit statistics are reported in the bottom section of **Table 21**. The estimated AIC (17,714) was higher than the BIC (17,410). The M_2 limited information goodness-of-fit statistic was 3,922, which was non-significant, and the RMSEA was 0.06. These results are difficult to compare across dimensions given different numbers of items. However, the non-significance of the M_2 statistic was as expected.

Table 21: Item calibrations - Physical functioning dimension

Item	*Slope*	*Item Thresholds*					*Item Level Fit*	
	α (se)[a]	*B_1 (se)*	*B_2 (se)*	*B_3 (se)*	*B_4 (se)*	*B Range*	*S-χ^{2f}*	*p*[g]
EQ-5D-5L MO	3.45 (0.24)	0.33 (0.05)	1.12 (0.06)	1.94 (0.10)	2.84 (0.20)	2.51	62.21	0.506
EQ-5D-5L SC	2.32 (0.21)	1.25 (0.07)	1.84 (0.11)	2.72 (0.19)	3.17 (0.26)	1.92	73.02	0.104
EQ-5D-5L UA	2.32 (0.16)	0.18 (0.05)	1.19 (0.07)	2.24 (0.13)	3.25 (0.26)	3.07	82.55	0.185
SF-36 Q3	2.01 (0.14)	-0.79 (0.07)	0.72 (0.06)	N/A	N/A	1.51	55.72	0.092
SF-36 Q4	2.91 (0.21)	0.23 (0.05)	1.46 (0.08)	N/A	N/A	1.23	70.29	0.012
SF-36 Q5	2.74 (0.20)	0.43 (0.05)	1.60 (0.09)	N/A	N/A	1.17	62.91	0.087
SF-36 Q6	3.02 (0.21)	-0.06 (0.05)	1.15 (0.06)	N/A	N/A	1.21	69.81	0.010
SF-36 Q7	4.24 (0.35)	0.54 (0.05)	1.51 (0.07)	N/A	N/A	0.97	65.72	0.006
SF-36 Q8	2.64 (0.18)	-0.08 (0.05)	1.28 (0.06)	N/A	N/A	1.36	56.39	0.189
SF-36 Q9	4.20 (0.32)	0.22 (0.05)	1.09 (0.06)	N/A	N/A	0.87	42.63	0.241
SF-36 Q10	4.67 (0.39)	0.55 (0.05)	1.28 (0.06)	N/A	N/A	0.73	44.87	0.175
SF-36 Q11	3.86 (0.34)	0.90 (0.05)	1.78 (0.09)	N/A	N/A	0.88	60.45	0.015
SF-36 Q12	2.53 (0.24)	1.24 (0.07)	2.22 (0.14)	N/A	N/A	0.98	87.40	<0.001
PROMIS Q1	2.49 (0.16)	-0.12 (0.05)	0.95 (0.06)	1.61 (0.09)	2.18 (0.12)	2.30	109.44	0.038
PROMIS Q2	2.97 (0.19)	0.05 (0.05)	0.85 (0.05)	1.41 (0.07)	1.98 (0.10)	1.93	99.56	0.079
PROMIS Q3	3.86 (0.28)	0.44 (0.05)	1.02 (0.06)	1.58 (0.08)	1.97 (0.10)	1.53	55.56	0.817
PROMIS Q4	2.84 (0.21)	0.64 (0.05)	1.30 (0.07)	1.95 (0.11)	2.49 (0.16)	1.85	91.18	0.038
Model fit statistics								
-2 * Log-likelihood	17,280							
AIC[b]	17,714							
BIC[c]	17,410							
M_2[d]	3,922							
RMSEA[e]	0.06							

[a] Standard Error; [b] Akaike Information Criterion; [c] Bayesian Information Criterion; [d] M_2 Goodness-of-Fit statistic; [e] Root Mean Squared Error of Approximation; [f] Chi Square item fit; [g] significance value

4.18.6. Assessing DIF

Table 22 displays the results of the DIF analysis, with bold values indicating significance. EQ-5D-5L self-care and usual activities were excluded from the gender DIF assessment due to lack of variation at the most severe response level, where all respondents reported the same gender. All three EQ-5D-5L items were excluded from the assessment of DIF by condition for the same reason. For the remaining items there was no significant DIF by gender. The four PROMIS items, and the SF-36 item assessing bathing and dressing displayed significant DIF by condition status,

where those with a condition scored significantly higher across the latent scale. This may be a sign of sensitivity of responses to the presence or absence of a long-term condition.

Table 22: DIF assessment by gender and condition – Physical functioning dimension

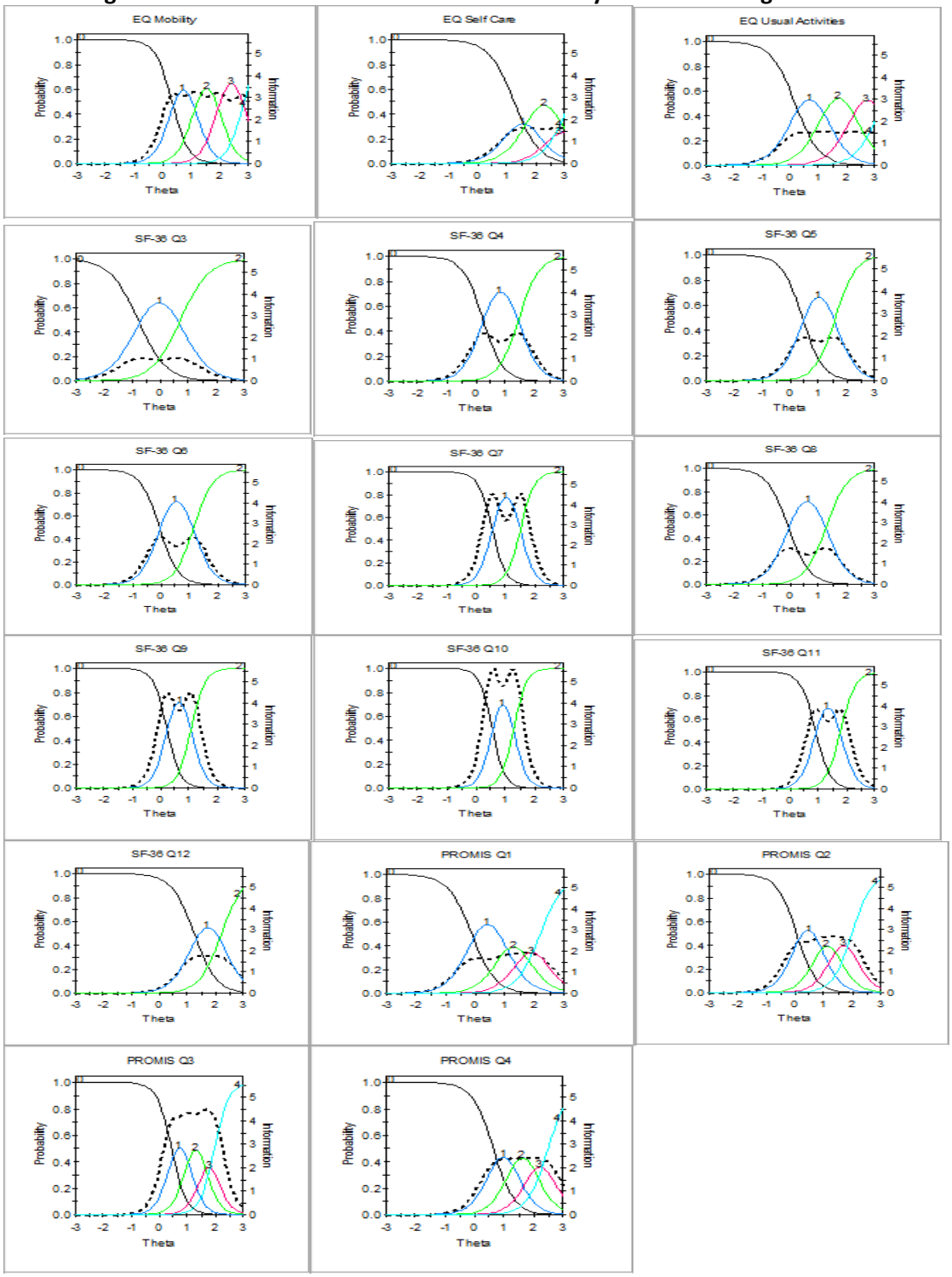
Item	Gender		Condition	
	X^{2a}	P^b	X^2	P
EQ MO	2.2	0.815	N/A	N/A
EQ SC	N/A	N/A	N/A	N/A
EQ UA	N/A	N/A	N/A	N/A
SF-36 Q3	2.0	0.582	**18.7**	**<0.001**
SF-36 Q4	3.9	0.269	0.2	0.976
SF-36 Q5	2.2	0.538	3.6	0.315
SF-36 Q6	1.7	0.645	3.2	0.361
SF-36 Q7	4.9	0.180	2.3	0.511
SF-36 Q8	1.1	0.772	8.3	0.040
SF-36 Q9	0.2	0.970	10.0	0.019
SF-36 Q10	1.0	0.809	0.2	0.979
SF-36 Q11	1.8	0.613	3.8	0.287
SF-36 Q12	0.1	0.987	**17.8**	**<0.001**
PROMIS Q1	1.9	0.861	**41.6**	**<0.001**
PROMIS Q2	1.7	0.886	**41.1**	**<0.001**
PROMIS Q3	3.6	0.603	**32.2**	**<0.001**
PROMIS Q4	2.2	0.818	**23.8**	**<0.001**

[a] Wald Chi Square; [b] Significance value

4.18.7. IRT item calibrations - Assessing item level thresholds and ordering

Table 21 reports the item calibrations, slopes and threshold parameters for the physical functioning dimension, and **Figure 13** reports the threshold curves. The threshold parameters demonstrated that the items displayed different response characteristics. For example, the threshold value for the transition between levels 1 and 2 on the EQ-5D-5L items occurred at different severity points on the theta scale (0.18, 0.33 and 1.25 for UA, MO and SC respectively, where a lower number indicates a transition at a lower severity of theta). Of the EQ-5D items, UA had the highest threshold for the transition between severe and extreme problems. The range of the threshold parameters for EQ UA indicated increased sensitivity to the underlying construct being measured. Across the other items, the item with the lowest transition point (-0.79, from 'not limited' to 'limited a little') was vigorous activity limitations. This was expected given that having any problems with vigorous activities was more likely to be reported than having problems across the other SF-36 items. The transition from being 'limited a little' to 'limited a lot' occurs at 0.72 which was before the first transition has occurred for other SF-36 items. For example, before the item assessing bathing and dressing limitations. The threshold transitions occurred at 1.24 and 2.22, which were both at a more severe point of the theta scale.

Figure 13: Threshold and information curves - Physical functioning dimension



The item with the largest overall coverage of theta was EQ-5D-5L usual activities (2.51) which occurred in the severe range of the theta. The item with the smallest range coverage was the SF-36v2 item assessing ability to walk several hundred metres (0.73). The SF-36v2 items generally covered a smaller overall severity range (linked to including less response levels), but

the ranges did appear at different points of theta. This suggests that they assess different severities of physical functioning.

Figure 13 demonstrates that, for most of the items, the threshold peaks appeared in order across the theta scale. This meant that the probability of the item responses appearing across the severity scale were ordered as expected. An exception to this was EQ-5D-5L SC, where threshold curves three and four did not appear as the highest peak demonstrating disordering for levels 4 (severe problems) and 5 (unable to).

4.18.8. IRT item calibrations - Assessing item information

Table 21 also reports the overall item slope parameters (α). The black dotted lines for each item on **Figure 13** demonstrate the information provided to the dimension by each item. The item slopes ranged from 2.01 to 4.67. The item curves suggested similar information profiles for a number of the items (for example EQ-5D-5L SC and UA). However, as they were sensitive over the central points of theta, they could both be sensitive indicators of a range of severities. **Table 23** and **Figure 14** report the total information provided by the item set. The overall dimension was most sensitive in the 0 to 1.6 (mild to moderate problems) range of theta. The standard error of measurement is inversely linked to the information provided, so was reduced as the information provided across theta increased. To improve the sensitivity of the dimension to less severe physical functioning problems, items assessing the mild problems may need to be developed to add to the information provided.

Table 23: Total test information at key points of the latent scale – Physical Functioning dimension

θ point[a]	Test information	Expected SE[b]
-2.4	1.22	0.90
-1.6	2.18	0.68
-0.8	7.05	0.38
0	27.33	0.19
0.8	43.33	0.15
1.6	40.83	0.16
2.4	19.47	0.23

[a] Selected severity points on underlying theta scale; [b] standard error

4.18.9. IRT Item calibrations - Assessing IRT score estimates

Figure 15 reports the expected score curve that can be used to link raw item total scores to the calibrated theta score. The total dimension score across the 17 items was 44 (as item response levels were coded from 0 to 4 for the EQ-5D-5L, 0 to 2 for the SF-36v2 items, and 0 to 3 for the

PROMIS-29). A theta score of zero was equivalent to a test score of 5 to 6, a theta score of one was equivalent to a test score of 20, and a theta of two was equivalent to a score of 37. This demonstrated the increase in sensitivity at the moderate range of theta, and suggested the need for more informative items for less severe problems.

Figure 14: Total information curve - Physical functioning dimension

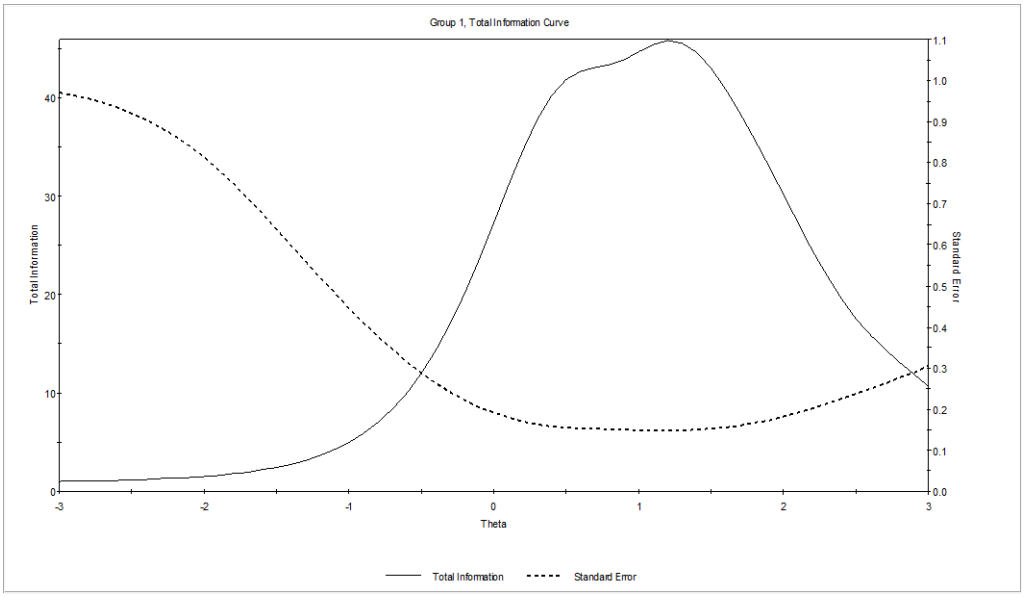
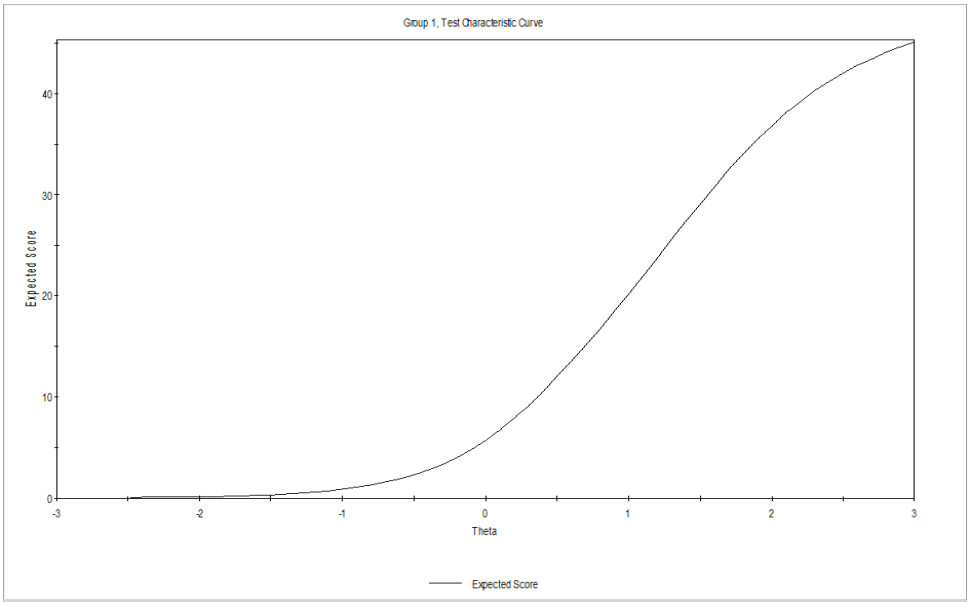


Figure 15: Expected score curve - Physical functioning dimension



4.18.10. Summary and implications of results

As noted above, the three functioning based EQ-5D-5L items (MO, SC and UA) loaded onto the same dimension. This has implications for the independence of the items. The IRT results also inform the relationship between the EQ-5D-5L items. The threshold analysis suggests similar

coverage of theta for MO and UA, with UA having a larger slope demonstrating that it provides more information. From a face validity perspective, it could be argued that the SF-36 and PROMIS items in this dimension are physical functioning focused, and therefore conceptually MO would be the preferred preference-based item for this dimension.

The different information profiles across the items included in the physical functioning dimension is also informative. The EQ-5D-5L items provide information for the mild to relatively severe level demonstrating that they cover the central range of theta, but may not be the most informative indicators for assessing mild problems (given the prevalence of answering 'no problems' for these items). The SF-36v2 items generally cover a smaller overall severity range which would be expected given they include less response levels. However, the ranges do appear at different points of the theta scale, and this indicates that they assess different severities of physical functioning. To improve the sensitivity of the dimension to less severe physical functioning problems, items assessing mild problems may need to be developed, or identified from existing measures.

4.19. Results – Mental health dimension

4.19.1. Justification of dimensionality

The dimensionality used for the mental health dimension included items that loaded with the mental health factor (with items from the EQ-5D-5L, SF-36 and PROMIS). This was used to explore the widest pool of mental health items as suggested by the EFA.

4.19.2. Initial data inspection

Table 24 reports the initial data inspection of the items. The response frequencies were spread across the five severity levels available for each item. This means that the respondent sample reported different severities of mental health problems, and the full response set was used in the analyses.

Table 24: Initial data inspection - Mental health dimension

Item	L1[a]	L2[b]	L3[c]	L4[d]	L5[e]
EQ-5D-5L AD	378 (47.6)	216 (27.2)	133 (16.8)	47 (5.9)	20 (2.5)
SF-36 Q24	290 (36.5)	211 (26.6)	188 (23.7)	83 (10.5)	22 (2.8)
SF-36 Q25	365 (46.0)	188 (23.7)	145 (18.3)	76 (9.6)	20 (2.5)
SF-36 Q26	65 (8.2)	253 (31.9)	234 (29.5)	185 (23.3)	57 (7.2)
SF-36 Q28	257 (32.4)	227 (28.6)	180 (22.7)	101 (12.7)	29 (3.7)
SF-36 Q30	75 (9.4)	304 (38.3)	241 (30.4)	134 (16.9)	40 (5.0)
PROMIS Q5	365 (46.0)	214 (27.0)	168 (21.2)	34 (4.3)	13 (1.6)
PROMIS Q6	385 (48.5)	177 (22.3)	172 (21.7)	49 (6.2)	11 (1.4)
PROMIS Q7	329 (41.4)	184 (23.2)	203 (25.6)	57 (7.2)	21 (2.6)
PROMIS Q8	252 (31.7)	245 (30.9)	213 (26.8)	62 (7.8)	22 (2.8)
PROMIS Q9	378 (47.6)	164 (20.7)	162 (20.4)	61 (7.7)	29 (3.7)
PROMIS Q10	355 (44.7)	191 (24.1)	163 (20.5)	61 (7.7)	24 (3.0)
PROMIS Q11	316 (39.8)	178 (22.4)	190 (23.9)	75 (9.4)	35 (4.4)
PROMIS Q12	384 (48.4)	168 (21.2)	150 (18.9)	68 (8.6)	24 (3.0)

[a] severity level 1 (least severe); [b] severity level 2; [c] severity level 3; [d] severity level 4; [e] severity level 5 (most severe)

4.19.3. Assessing local dependence

Overall 20 (25.6%) of pairs displayed potential local dependence (with a score above 10). **Table 25** reports the 20 item pairs in detail, and Appendix 10 reports all item pair Chi Square estimates. The SF-36v2 item assessing frequency of 'been happy' displayed local dependence across nine item pairs, including items from both the SF-36v2 and the PROMIS-29. To refine the dimension, this item could be considered for exclusion. The SF-36v2 item assessing the frequency of feeling 'calm and peaceful' appeared across seven item pairs (including with the item about 'been happy', which has the second largest item pair score).

Table 25: Item pairs with local dependence estimates > 10 - Mental health dimension

Item one	Item two	X^{2a}
SF-36 28: Downhearted and depressed	SF-36 30: Been happy	28.1
SF-36 26: Calm and peaceful	SF-36 30: Been happy	25.7
SF-36 26: Calm and peaceful	PROMIS 8: Felt uneasy	22.9
SF-36 25: Down in the dumps	SF-36 30: Been happy	18.8
SF-36 25: Down in the dumps	SF-36 26: Calm and peaceful	18.7
SF-36 26: Calm and peaceful	SF-36 28: Downhearted and depressed	17.6
SF-36 30: Been happy	PROMIS 8: Felt uneasy	15.7
SF-36 26: Calm and peaceful	PROMIS 6: Hard to focus on anything other than anxiety	13.4
SF-36 24: Very nervous	SF-36 26: Calm and peaceful	13.3
PROMIS 5: I felt fearful	PROMIS 9: I felt worthless	12.6
SF-36 28: Downhearted and depressed	PROMIS 11: I felt depressed	12.6
SF-36 24: Very nervous	SF-36 30: Been happy	12.0
SF-36 30: Been happy	PROMIS 5: Felt fearful	12.0
SF-36 30: Been happy	PROMIS 7: Worries overwhelmed me	12.0
SF-36 30: Been happy	PROMIS 11: I felt depressed	11.9

PROMIS 6: Hard to focus on anything other than anxiety	PROMIS 9: I felt worthless	11.9
SF-36 26: Calm and peaceful	PROMIS 10: Felt helpless	11.5
SF-36 25: Down in the dumps	SF-36 28: Downhearted and depressed	11.5
SF-36 26: Calm and peaceful	PROMIS 11: I felt depressed	10.7
SF-36 30: Been happy	PROMIS 10: Felt helpless	10.5

[a] Standardised Chi Square

4.19.4. Assessing model-data fit – Item level

Table 26 reports the results of the item level fit statistics. Three items (SF-36v2 items 'calm and peaceful' and 'been happy', and the PROMIS-29 item assessing the frequency of 'feeling fearful') displayed significant misfit to the model, so could be iteratively considered for removal based on other evidence such as the results of the local dependency analysis.

4.19.5. Assessing model-data fit – Model level

Table 26 also reports the model level fit statistics. In contrast to the physical functioning dimension, the BIC (20,235) was larger than the AIC (19,942). The M_2 goodness-of-fit statistic was 8,308 which was higher than the equivalent for the physical functioning dimension and was non-significant as expected. The RMSEA is 0.08 which was higher than the physical functioning dimension.

Table 26: Item calibrations – Mental health dimension

Item	***Slope***	***Item Thresholds***					***Item Level Fit***	
	α (se)[a]	***B_1 (se)***	***B_2 (se)***	***B_3 (se)***	***B_4 (se)***	***B Range***	***S-χ^{2f}***	***$p^{g.}$***
EQ-5D-5L AD	3.13 (0.19)	-0.05 (0.05)	0.75 (0.05)	1.54 (0.07)	2.23 (0.12)	2.28	92.52	0.142
SF-36 Q24	2.20 (0.13)	-0.44 (0.06)	0.43 (0.05)	1.41 (0.08)	2.45 (0.14)	2.89	98.11	0.365
SF-36 Q25	3.15 (0.19)	-0.10 (0.05)	0.63 (0.05)	1.35 (0.07)	2.22 (0.12)	2.32	101.42	0.062
SF-36 Q26	1.72 (0.11)	-2.03 (0.13)	-0.33 (0.06)	0.77 (0.07)	2.10 (0.13)	4.13	204.44	<0.001
SF-36 Q28	3.34 (0.19)	-0.50 (0.05)	0.34 (0.05)	1.10 (0.06)	2.00 (0.10)	2.50	87.98	0.229
SF-36 Q30	1.85 (0.11)	-1.83 (0.11)	-0.06 (0.06)	1.08 (0.07)	2.28 (0.13)	4.11	166.51	<0.001
PROMIS Q5	2.65 (0.16)	-0.12 (0.05)	0.70 (0.05)	1.79 (0.09)	2.55 (0.15)	2.67	135.59	<0.001
PROMIS Q6	3.94 (0.25)	-0.03 (0.05)	0.61 (0.05)	1.53 (0.07)	2.36 (0.12)	2.39	71.13	0.341
PROMIS Q7	3.64 (0.22)	-0.22 (0.05)	0.45 (0.05)	1.41 (0.07)	2.12 (0.10)	2.34	86.71	0.130
PROMIS Q8	3.38 (0.20)	-0.52 (0.05)	0.38 (0.05)	1.38 (0.07)	2.14 (0.11)	2.66	95.77	0.045
PROMIS Q9	3.92 (0.24)	-0.05 (0.05)	0.55 (0.05)	1.30 (0.06)	1.93 (0.09)	1.98	97.31	0.030
PROMIS Q10	3.88 (0.24)	-0.11 (0.05)	0.55 (0.05)	1.32 (0.06)	2.05 (0.10)	2.16	93.03	0.041
PROMIS Q11	4.67 (0.30)	-0.26 (0.05)	0.36 (0.04)	1.14 (0.05)	1.79 (0.08)	2.05	68.33	0.332
Model fit statistics								
-2 * Log-Likelihood	19,812							
AIC[b]	19,942							
BIC[c]	20,246							
M2[d]	8,308							
RMSEA[e]	0.08							

[a] Standard Error; [b] Akaike Information Criterion; [c] Bayesian Information Criterion; [d] M_2 Goodness-of-Fit statistic; [e] Root Mean Squared Error of Approximation; [f] Chi Square item fit; [g] significance value

4.19.6. Assessing DIF

Table 27 reports the results of the DIF analysis by gender and condition, with the items displaying significant DIF in bold. There was no significant evidence of DIF by gender for any item at the overall level. There was evidence of DIF by condition status at $p < 0.001$ for three SF-36v2 items (feeling 'down in the dumps', 'calm and peaceful' and 'happy'). This indicates that having a long-term condition results in different responses to these items.

Table 27: DIF by gender and condition - Mental health dimension

Item	Gender		Condition	
	X^{2a}	P^b	X^2	p
EQ AD	1.2	0.941	**19.4**	**0.002**
SF-36 Q24	4.5	0.484	9.5	0.091
SF-36 Q25	2.7	0.742	**23.2**	**<0.001**
SF-36 Q26	4.9	0.423	**22.8**	**<0.001**
SF-36 Q28	1.2	0.949	**14.9**	**0.011**
SF-36 Q30	3.8	0.583	**26.9**	**<0.001**
PROMIS Q5	2.0	0.853	5.2	0.389
PROMIS Q6	4.1	0.530	8.5	0.130
PROMIS Q7	3.5	0.622	6.9	0.230
PROMIS Q8	3.1	0.679	1.7	0.890
PROMIS Q9	1.0	0.961	11.9	0.037
PROMIS Q10	5.1	0.400	9.0	0.108
PROMIS Q11	2.7	0.744	6.7	0.248

4.19.7. IRT item calibrations - Assessing item level thresholds and ordering

Table 26 reports the item calibrations for the mental health dimension, and **Figure 16** reports the threshold curves. The EQ-5D-5L AD item was sensitive across the mild to moderate range of theta. The items with the lowest threshold transition were the SF-36v2 items assessing been 'clam and peaceful' and 'happy', at -2.03 and -1.83 respectively. This meant that at more severe points of underlying mental health scale, respondents will state that they are calm and peaceful or happy 'a little of the time' instead of 'none of the time'. These two items displayed the largest overall coverage of theta (4.13 and 4.11 respectively), meaning that the transition from been 'calm and peaceful' and 'happy' 'most of the time' to 'all of the time' occurs at mild overall severity. Most of the other SF-36v2 and PROMIS-29 items covered a similar range of mild to moderate severities. **Figure 16** demonstrates that all item response curves were ordered indicating that the response levels were operating as expected, and the severity of the levels could be distinguished by respondents with different levels of mental health concerns.

4.19.8. IRT item calibrations - Assessing item information

Table 26 reports the item slopes as an indicator of the information provided by each item. The range of slope values ranged from 1.72 (SF-36v2 calm and peaceful) to 4.67 (PROMIS-29 feeling depressed), with the majority of items above three. This demonstrated that each item provided substantial information to the overall dimension. The EQ-5D-5L anxiety/depression item slope was 3.13, which, linked to the threshold scores, demonstrated that it was sensitive to the middle range of the scale and could therefore to act as a general indicator of the severity of mental health as measured by the overall theta scale.

Figure 16 reports the information characteristic curves for each item (the dotted lines). The information profiles reflect the slope values, but their characteristics across the theta scale clearly differed. For example, the information curve for EQ-5D-5L AD peaked across the mild to moderate theta scale, where theta ranges from approximately 0 to 2. In comparison, SF-36v2 items 26 (calm and peaceful) and 30 (happy) provided a low level of information across the overall theta scale. The majority of the PROMIS-29 items had different profiles, where they provided information across the mild to moderate theta range, with a reduction in the information provided at the central response category threshold points.

Figure 16: Threshold and information curves - Mental health dimension

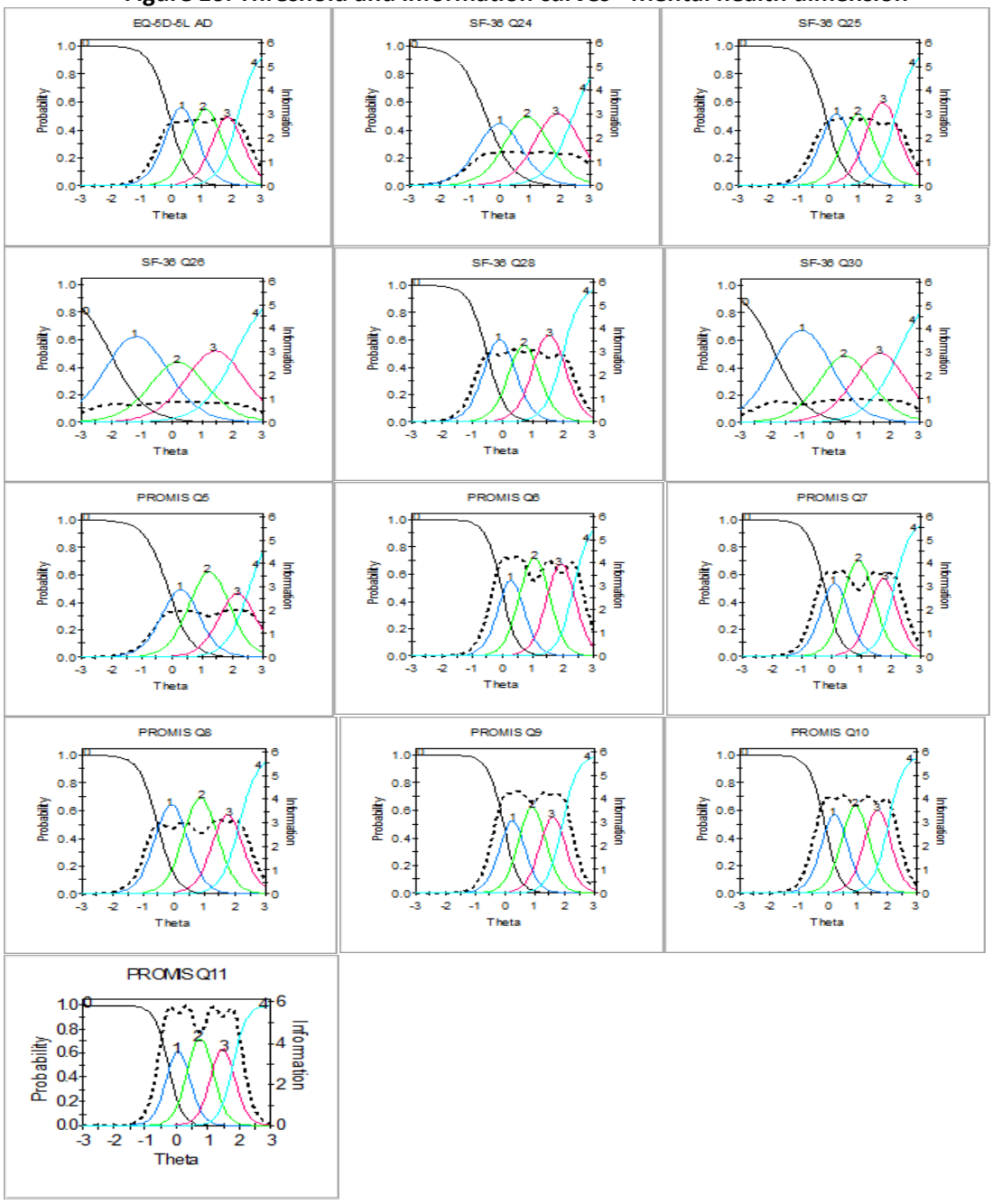


Figure *17* displays the information curve for the overall dimension, and **Table 28** reports the information provided at key points of the latent scale. As with the physical functioning dimension, there was less information provided, and therefore lower sensitivity, at the less severe range of theta. The standard error followed the same inverse pattern as the physical functioning dimension. The information provided, and the sensitivity to mental health issues,

peaked around the points where theta ranged from 0 and 1.6. This demonstrated measurement sensitivity to a wide range of mild to moderate problems.

Table 28: Total test information at key points of the latent scale - Mental Health dimension

θ point[a]	Test information	Expected SE[b]
-2.4	2.48	0.64
-1.6	4.06	0.50
-0.8	15.80	0.25
0	38.78	0.16
0.8	36.40	0.17
1.6	38.56	0.16
2.4	26.83	0.19

[a] Selected severity points on underlying theta scale; [b] standard error

4.19.9. IRT item calibrations - Assessing IRT score estimates

Figure 18 reports the estimated score curve for the mental health dimension. The total raw score based on the items was 52. A theta of zero was equivalent to a score of 12 to 13, a theta of one was equivalent to a score of 27, and a theta of two was equivalent to scoring 42. As with physical functioning, an exponential increase was observed, with relatively low sensitivity at the mild end of the scale demonstrated.

Figure 17: Total information curve - Mental health dimension

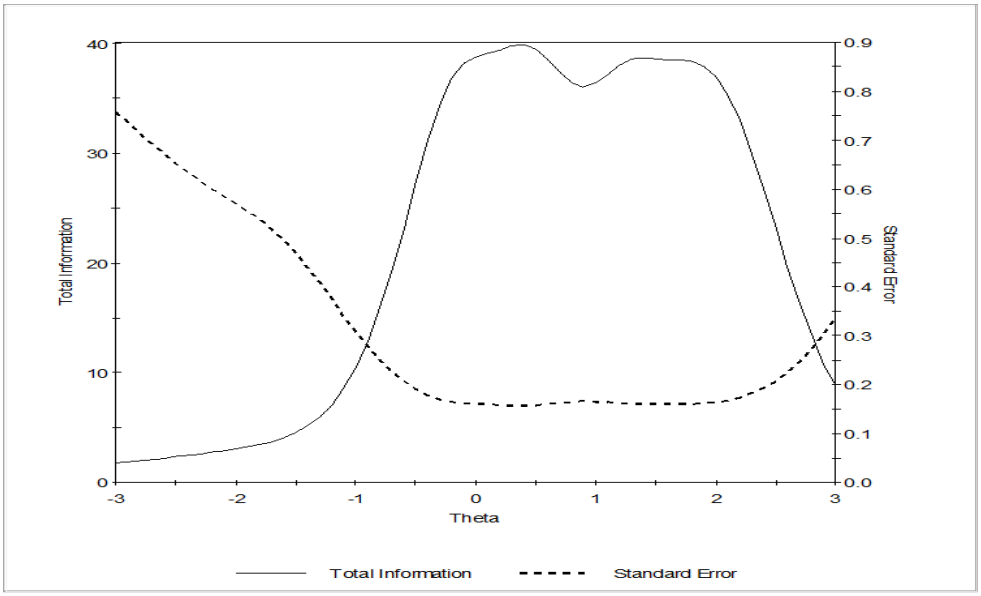
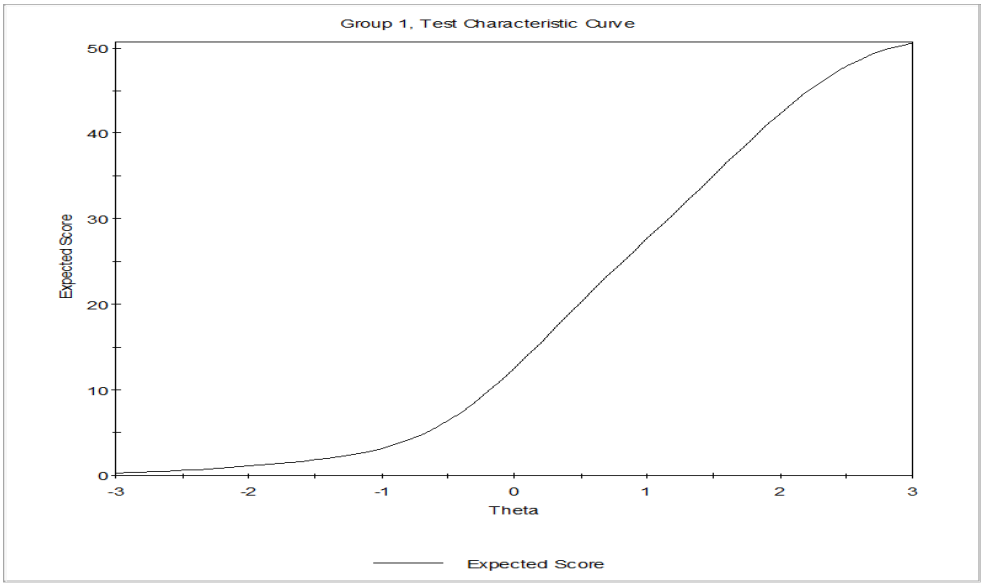


Figure 18: Estimated score curve - Mental health dimension



4.19.10. Summary and implications of results

For this dimension, the EQ-5D AD item could be used as the preference-based layer, as it was demonstrated to be sensitive across a wide range of theta. This means that the item may be acceptable indicator of general mental health issues from to which other more specific items could be calibrated for the collection of more detailed information about specific areas of broader mental health issues. This is further demonstrated by the results for the SF-36 and PROMIS items which have different characteristics across the severity scale. This could be because they ask about specific facets of broader mental health with both positive and negative framing of the items (see Appendix 9 for full detail of the item wording).

However, the IRT results do suggest some issues with the measurement of mental health. At the overall level, there was less information provided, and therefore lower sensitivity, at the mild range of theta. This is a result of the item content, and as with physical functioning, items could be developed or sourced from other existing measures, to increase measurement sensitivity to milder mental health issues. Linked to this, the dimension does cover more severe mental health concerns, but is focused on commonly occurring issues such as anxiety and depression rather than more severe issues such as psychosis or schizophrenia (where it has been shown that generic instruments are not psychometrically valid [114]). This demonstrates the limitations of using a layered approach based on existing instruments. For more severe mental health concerns, alternative measures would be required.

There is some evidence of local dependence which suggests that combining items from different instruments can lead to overlap the concepts that are been measured, and the information provided by these items. Extending the item pool to include a broader range of measures could extend the concepts included within dimensions where local dependence issues may require the removal of overlapping items.

4.20. Results – Pain dimension

4.20.1. Justification of dimensionality

Across both EFA models, the dimension defined as measuring pain included the items from the EQ-5D, SF-36 and PROMIS-29 directly specified as assessing pain. This provided the basis for including all of the items in the pain dimension.

4.20.2. Initial data inspection

Table 29 reports the initial data inspection for the pain dimension. Responses were distributed across each level of each item. This indicated that the items were sensitive to differing pain levels within the sample, and the data was acceptable for the analysis undertaken.

Table 29: Initial data description - Pain dimension

Item	L1[a]	L2[b]	L3[c]	L4[d]	L5[e]	L6[f]
EQ-5D-5L PD	238 (30.0)	336 (42.3)	151 (19.0)	55 (6.9)	14 (1.8)	N/A
SF-36 Q21	122 (15.4)	231 (29.1)	170 (21.4)	198 (24.9)	54 (6.8)	19 (2.4)
SF-36 Q22	295 (37.2)	259 (32.6)	143 (18.0)	77 (9.7)	20 (2.5)	N/A
PROMIS Q1	361 (45.5)	259 (32.6)	100 (12.6)	44 (5.5)	30 (3.8)	N/A
PROMIS Q25	296 (37.3)	270 (34.0)	129 (16.2)	64 (8.1)	35 (4.4)	N/A
PROMIS Q26	315 (39.7)	253 (31.9)	128 (16.1)	62 (7.8)	36 (4.5)	N/A
PROMIS Q27	430 (54.2)	172 (21.7)	120 (15.1)	41 (5.2)	31 (3.9)	N/A
PROMIS Q28	357 (45.0)	228 (28.7)	117 (14.7)	65 (8.2)	27 (3.4)	N/A

[a] severity level 1 (least severe); [b] severity level 2; [c] severity level 3; [d] severity level 4; [e] severity level 5; [f] severity level 6 (most severe)

4.20.3. Assessing local dependence

Table 30 provides detail of the item pairs with a score greater than 10, and Appendix 10 reports the results of the local dependence analysis for all items included in the dimension. Overall, 5 item pairs (17.9% of the 28 possible combinations) displayed local dependence. All five included the EQ-5D-5L PD item. The item pair with the highest Chi Square value provided an informative example of dependence. This was EQ-5D-5L PD and the SF-36v2 item assessing pain severity.

The EQ-5D item asks about pain and discomfort on a five level severity scale (none, slight, moderate, severe and extreme). The SF-36v2 item asks about bodily pain on a six point scale (none, very mild, mild, moderate, severe and very severe). Therefore, clear similarities in item content and response levels can be observed.

Table 30: Item pairs with local dependence estimates > 10 - Pain dimension

Item one	Item two	X^{2a}
EQ Pain/Discomfort	SF-36 21: Pain severity	18.2
EQ Pain/Discomfort	PROMIS 25: pain interfere with your day to day activities?	12.6
EQ Pain/Discomfort	PROMIS 26: pain interfere with work around the home	10.8
EQ Pain/Discomfort	PROMIS 27: pain interfere with ability to participate in social activities	11.0
EQ Pain/Discomfort	PROMIS 28: Pain interfere with household chores	11.4

[a] Standardised Chi Square

4.20.4. Assessing model-data fit – Item level

Table 31 reports the item level fit statistics. The majority of the items displayed evidence of misfit indicating differences in what was measured by the items in this dimension. To refine the dimension, items could be iteratively removed to improve fit. The sequence of removing items could be linked to those with the highest level of misfit, or combining information from other indicators tested.

4.20.5. Assessing model-data fit – Model level

Table 31 also reports the model level fit statistics. In line with the mental health dimension, the BIC (11,710) was larger than the AIC (11,518). The M_2 goodness-of-fit statistic was 2,603 (which was lower than the previous analyses, but the dimension included less items), and the RMSEA was 0.08.

Table 31: Item calibrations - Pain dimension

Item	*Slope*	*Item thresholds*						*Item level fit*	
	α (se)[a]	*B_1 (se)*	*B_2 (se)*	*B_3 (se)*	*B_4 (se)*	*B_5 (se)*	*B range*	*S-χ^{2f}*	*p*[g]
EQ-5D-5L PD	2.55 (0.15)	-0.64 (0.07)	0.67 (0.06)	1.61 (0.09)	2.59 (0.16)	N/A	3.23	183.55	<0.001
SF-36 Q21	2.97 (0.17)	-1.20 (0.08)	-0.14 (0.06)	0.47 (0.06)	1.51 (0.08)	2.31 (0.13)	3.51	99.31	0.006
SF-36 Q22	3.32 (0.20)	-0.35 (0.06)	0.59 (0.06)	1.28 (0.07)	2.20 (0.12)	N/A	2.55	68.57	0.121
PROMIS Q1	2.02 (0.13)	-0.14 (0.07)	1.01 (0.08)	1.73 (0.10)	2.36 (0.14)	N/A	2.50	111.59	0.009
PROMIS Q25	6.79 (0.51)	-0.31 (0.06)	0.59 (0.05)	1.14 (0.06)	1.70 (0.08)	N/A	2.01	73.98	<0.001
PROMIS Q26	8.89 (0.94)	-0.23 (0.06)	0.59 (0.05)	0.13 (0.06)	1.65 (0.08)	N/A	1.88	82.76	<0.001
PROMIS Q27	4.22 (0.28)	0.13 (0.06)	0.74 (0.06)	1.38 (0.07)	1.86 (0.09)	N/A	1.73	79.85	0.004
PROMIS Q28	6.67 (0.51)	-0.09 (0.06)	0.66 (0.05)	1.18 (0.06)	1.82 (0.08)	N/A	1.91	76.14	<0.001
Model fit statistics									
-2 * LL	11,436								
AIC[b]	11,518								
BIC[c]	11,710								
M2[d]	2,603								
RMSEA[e]	0.08								

[a] Standard Error; [b] Akaike Information Criterion; [c] Bayesian Information Criterion; [d] M_2 Goodness-of-Fit statistic; [e] Root Mean Squared Error of Approximation; [f] Chi Square item fit; [g] significance value

4.20.6. Assessing DIF

Table 32 reports DIF by gender and condition for each pain dimension item. There was significant evidence of DIF by gender ($p < 0.001$) for three of the PROMIS-29 items, with the item assessing how much pain interfered with work around the home the most impacted (female respondents provided significantly higher responses across the theta scale). Female respondents also provided higher responses on the other two items (pain interfering with day to day activities, and social activities). This is an indication that the items are sensitive to between group characteristics. There was also significant ($p < 0.001$) evidence for DIF by condition for the PROMIS item assessing ability to do chores.

Table 32: DIF by gender and condition - Pain dimension

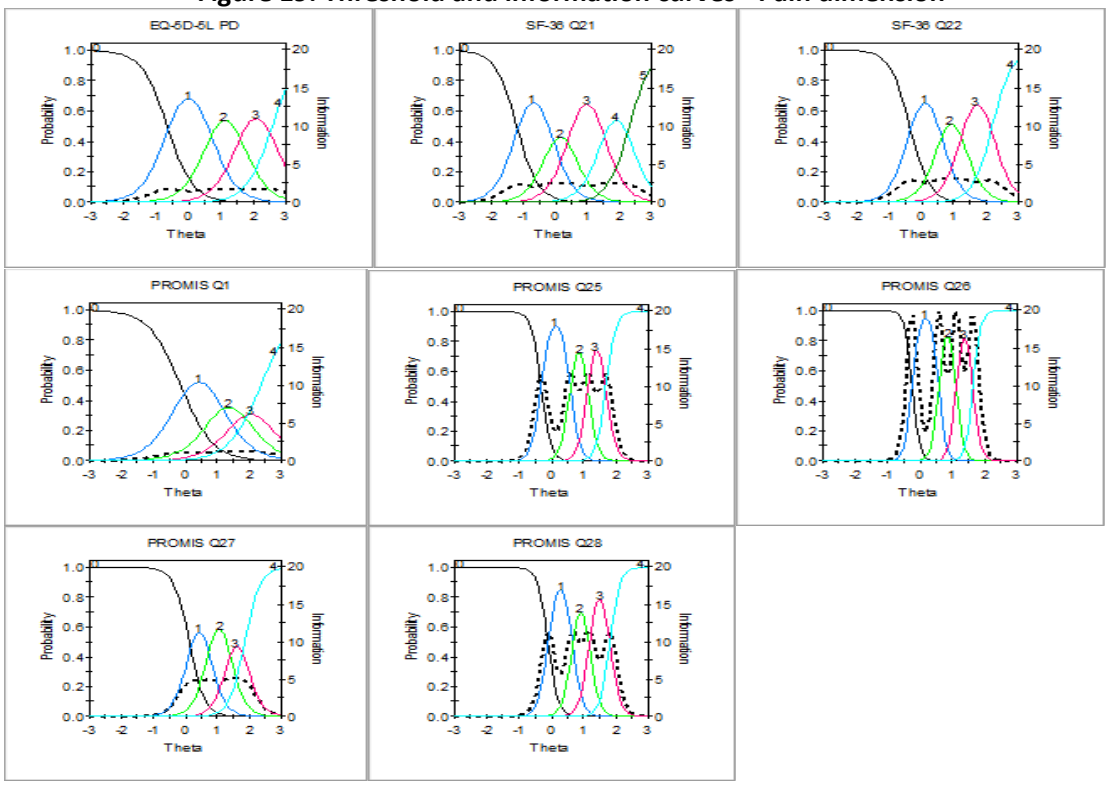
Item	Gender		Condition	
	X^{2a}	P[b]	X^2	P
EQ PD	10.6	0.061	N/A	N/A
SF-36 Q21	12.1	0.060	14.9	0.021
SF-36 Q22	10.6	0.061	11.1	0.050
PROMIS Q1	12.7	0.026	**30.2**	**<0.001**
PROMIS Q25	**20.5**	**0.001**	1.3	0.933
PROMIS Q26	**295.7**	**<0.001**	13.0	0.024
PROMIS Q27	**20.9**	**<0.001**	4.3	0.508
PROMIS Q28	12.6	0.028	2.7	0.748

[a] Wald Chi Square; [b] Significance value

4.20.7. IRT item calibrations - Assessing item level thresholds and ordering

Table 31 reports the item calibrations for the pain dimension. The SF-36v2 pain severity item had the lowest threshold parameter at -1.20 (between no pain and very mild pain), and this was followed by the EQ-5D-5L PD where the threshold between no problems and slight problems was -0.64. These two items also covered the largest range of theta, at 3.51 and 3.23 respectively, with the PROMIS items covering a smaller overall range, but four of the five items included a negative parameter value which indicated that they had sensitivity to milder pain problems to some extent. The wider coverage of the pain severity items was expected given the general nature of the question wording. **Figure 19** demonstrates that the response levels for each item were operating as expected, as the probability of each level appearing was ordered as the severity of theta increased.

Figure 19: Threshold and information curves - Pain dimension



4.20.8. IRT item calibrations - Assessing item information

Table 31 reports the slopes and information curved for each item. The PROMIS-29 items that directly measuring pain impacts (25 to 28) had the largest slopes which indicated an increased amount of information across the shorter range of theta where the item thresholds were operating. The profile of the information curves for items 25, 26 and 28 peaked at the threshold parameter valued, but decreased at the points of theta where no transition occurred. In contrast the curve for item 27 was smoother across theta. Comparing the PROMIS curve profiles with the EQ-5D-5L PD and SF-36v2 items was instructive. The pain severity items provided a consistent but low level of information. This indicated their usefulness as general indicators of pain, where the PROMIS-29 items added to this general information and provided more detail across a narrower range. **Figure 20** reports the overall information curve, and **Table 33** reports the total information at different thetas. The combined curve was impacted by the inconsistent profile of the PROMIS-29 items, and was reduced at certain points of theta. Again, the standard error reduced as the information provided increased.

Table 33: Total test information at key points of the latent scale - Pain dimension

θ point[a]	Test information	Expected SE[b]
-2.4	1.36	0.86
-1.6	3.45	0.54
-0.8	9.76	0.32
0	36.87	0.16
0.8	46.47	0.15
1.6	53.06	0.14
2.4	11.58	0.29

[a] Selected severity points on underlying theta scale; [b] standard error

4.20.9. IRT item calibrations - Assessing IRT score estimates

Figure 21 displays the estimated score curves for the pain dimension, with a total score of 33 possible. A theta of zero was equivalent to a score of 7, one was equivalent to a score of 17, and two was equivalent to scoring 28. This was a similar theta profile across the score range to the physical functioning and mental health dimensions.

Figure 20: Total information curve - Pain dimension

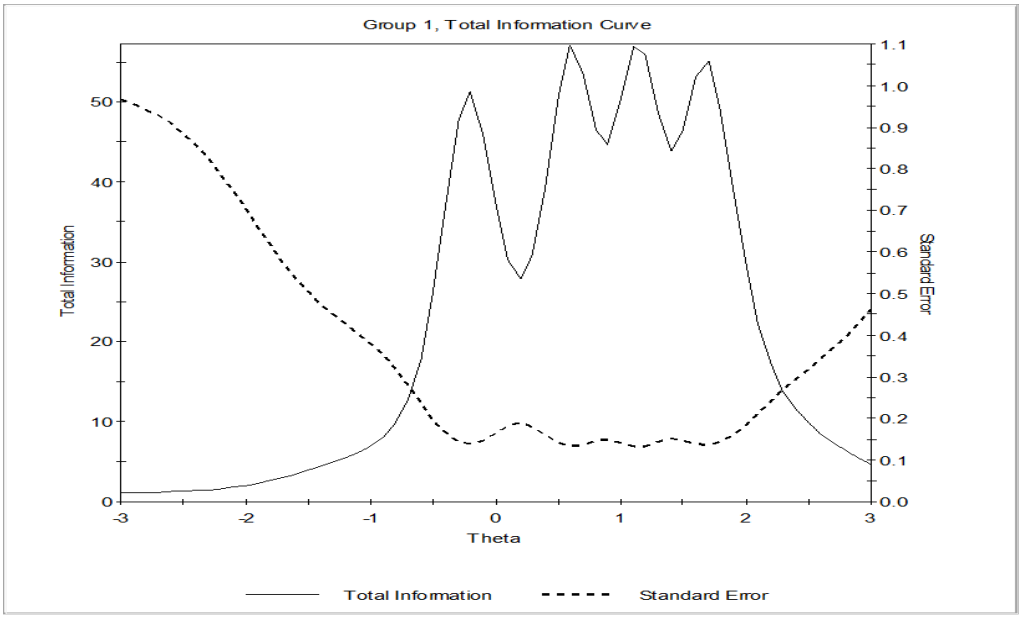
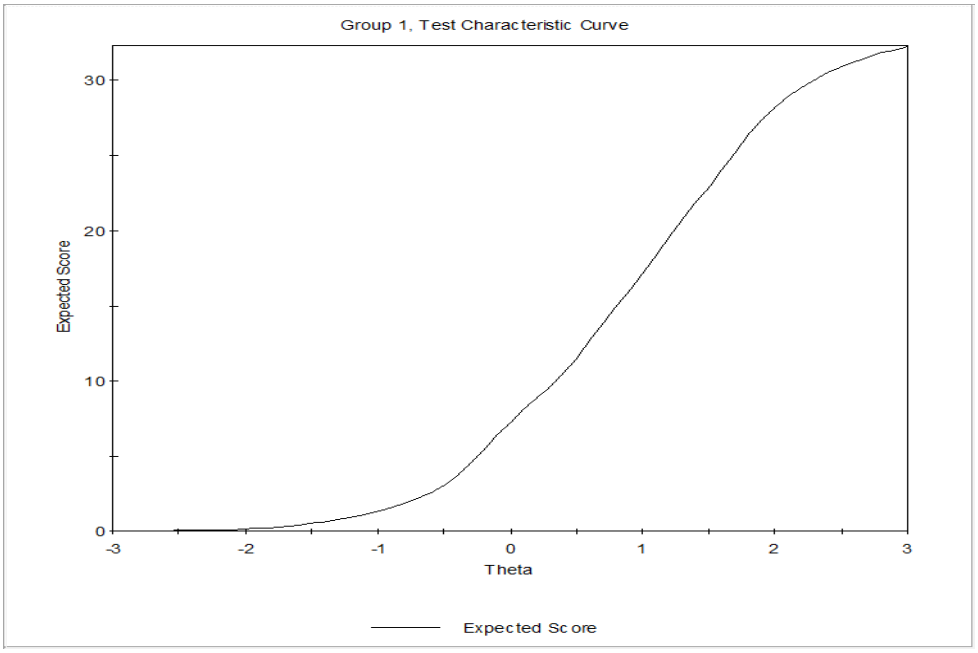


Figure 21: Estimated score curve - Pain dimension



4.20.10. Summary and implications of results

The analysis of the dimension defined as pain demonstrates that the EQ-5D and SF-36 items that assess pain severity are useful general indicators that provide a low level of information across the theta scale. This is quite different to the information characteristics of the PROMIS-29 items which provide increased information at certain points of theta translating to the threshold points. A number of the items displayed misfit to the underlying model, which could be due to differences in the pain related concepts measured (for example differences in pain interference and impacts, and pain severity). That the PROMIS-29 items displayed misfit, and also evidence of DIF, is of interest given these items were originally developed using an IRT approach. However, this could be due to combining diverse items, and also be a result of the sample used. The total information provided is also inconsistent across the severity scale, and developing or adding items to test improving the consistency of the dimension could support the wider development of a layered approach to measuring pain.

4.21. Results – Activities dimension

4.21.1. Justification of dimensionality

The items included in the activities dimension did not have a factor loading above 0.4 in the EFA models. However, as the non-loading items assessed a similar underlying construct based on the

source measures, and fitted with the EQ-5D-5L UA dimension, the performance of the items was tested to understand their relationship and fit.

4.21.2. Initial data inspection

Table 34 reports the frequency of respondents who answered at each level of each item. The responses were spread across the available levels, again indicating that the items had a level of sensitivity to issues with activities reported by the sample, and the data was acceptable for IRT analyses.

Table 34: Initial data inspection – Activities dimension

Item	L1[a]	L2[b]	L3[c]	L4[d]	L5[e]
EQ-5D-5L UA	439 (55.3)	224 (28.2)	102 (12.8)	24 (3.0)	5 (0.6)
SF-36 Q20	336 (42.3)	208 (26.2)	150 (18.9)	69 (8.7)	31 (3.9)
SF-36 Q32	323 (40.7)	197 (24.8)	160 (20.2)	79 (9.9)	35 (4.4)
PROMIS Q21	258 (32.5)	241 (30.4)	197 (24.8)	64 (8.1)	34 (4.3)
PROMIS Q22	266 (33.5)	224 (28.2)	203 (25.6)	71 (12.7)	30 (4.0)
PROMIS Q23	241 (30.4)	231 (29.1)	217 (27.3)	77 (9.7)	28 (3.5)
PROMIS Q24	276 (34.8)	206 (25.9)	19 (24.6)	72 (9.1)	45 (5.7)

[a] severity level 1 (least severe); [b] severity level 2; [c] severity level 3; [d] severity level 4; [e] severity level 5 (most severe)

4.21.3. Assessing local dependence

Table 35 and Appendix 10 report the results of the local dependence analysis for the activities dimension. Overall, two (9.5%) of the item pairs displayed differing levels of local dependence. The item pair with a high level of dependence included two SF-36v2 items assessing the extent, and frequency, that physical health and emotional problems interfered with social activities. The EQ-5D-5L UA item did not display evidence of local dependence with the other items.

Table 35: Item pairs with local dependence estimates > 10 - Activities dimension

Item one	Item two	X^{2a}
SF-36 20: Extent physical health or emotional problems interfered with social activities	SF-36 32: Frequency physical health or emotional problems interfered with social activities	25.8
PROMIS 23: Trouble doing all of usual work	PROMIS 24: Trouble doing all of the activities with friends	11.4

[a] Standardised Chi Square

4.21.4. Assessing model-data fit – Item level

Table 36 reports the item level fit statistics. Two PROMIS-29 items did not fit the model ($p < 0.01$), with relatively low fit for the other PROMIS-29 and SF-36v2 items. The EQ-5D-5L UA item was not a significant outlier.

4.21.5. Assessing model-data fit – Model level

Table 36 also includes the model level fit statistics. The BIC was larger than the AIC, which was the case with the mental health and pain dimensions. The M2 goodness-of-fit statistic was 1,422 (lower than the other dimensions due to item numbers), and the RMSEA was 0.07.

Table 36: Item calibrations - Activities dimension

Item	*Slope*	*Item thresholds*					*Item level fit*	
	α (se)[a]	*B_1 (se)*	*B_2 (se)*	*B_3 (se)*	*B_4 (se)*	*B range*	*S-χ^{2f}*	*p*[g]
EQ-5D-5L UA	2.08 (0.14)	0.17 (0.06)	1.23 (0.08)	2.35 (0.14)	3.38 (0.27)	3.21	46.43	0.692
SF-36 Q20	1.98 (0.13)	-0.23 (0.06)	0.66 (0.06)	1.53 (0.09)	2.39 (0.15)	2.62	88.61	0.018
SF-36 Q32	2.12 (0.13)	-0.28 (0.06)	0.54 (0.06)	1.38 (0.08)	2.24 (0.13)	2.52	85.45	0.031
PROMIS Q21	4.25 (0.27)	-0.47 (0.05)	0.37 (0.05)	1.24 (0.06)	1.85 (0.09)	2.32	73.41	0.003
PROMIS Q22	4.98 (0.35)	-0.43 (0.05)	0.33 (0.05)	1.18 (0.06)	1.78 (0.09)	2.21	61.07	0.002
PROMIS Q23	3.53 (0.21)	-0.55 (0.06)	0.28 (0.05)	1.20 (0.06)	2.00 (0.10)	2.55	68.80	0.026
PROMIS Q24	5.54 (0.41)	-0.38 (0.05)	0.30 (0.04)	1.09 (0.06)	1.63 (0.07)	2.01	58.17	0.019
Model fit statistics								
-2 * Log-Likelihood	11,070							
AIC[b]	11,138							
BIC[c]	11,297							
M2[d]	1,422							
RMSEA[e]	0.07							

[a] Standard Error; [b] Akaike Information Criterion; [c] Bayesian Information Criterion; [d] M_2 Goodness-of-Fit statistic; [e] Root Mean Squared Error of Approximation; [f] Chi Square item fit; [g] significance value

4.21.6. Assessing DIF

There was no significant DIF by gender or condition across any of the items. The EQ-5D-5L UA dimension was excluded from the gender analysis given lack of gender differences between the respondents at the most severe level. The DIF results are reported in **Table 37**.

Table 37: DIF by gender and condition - Activities dimension

Item	Gender		Condition	
	X^{2a}	P[b]	X2	P
EQ-5D-5L UA	N/A	N/A	6.1	0.295
SF-36 20	3.8	0.582	7.3	0.201
SF-36 32	5.5	0.360	6.1	0.294
PROMIS 21	3.2	0.667	7.2	0.209
PROMIS 22	0.4	0.984	1.4	0.846
PROMIS 23	2.5	0.780	1.4	0.920
PROMIS 24	3.4	0.644	2.8	0.737

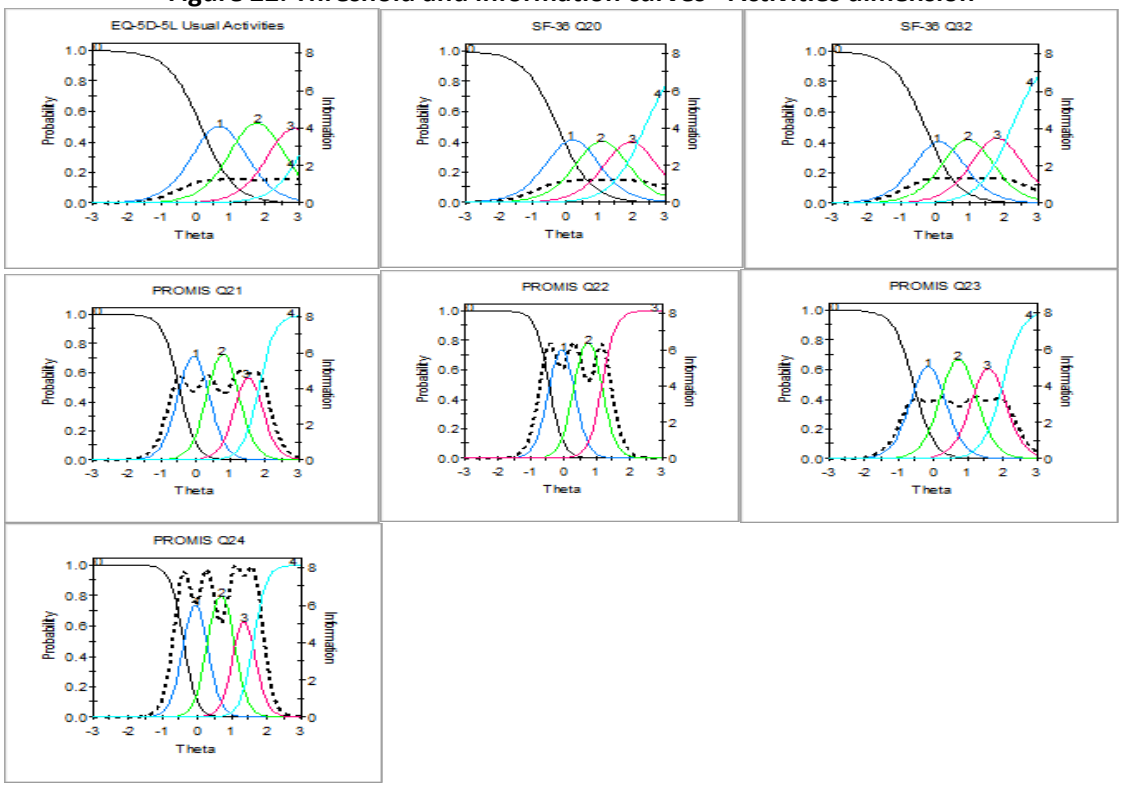
[a] Wald Chi Square; [b] Significance value

4.21.7. IRT item calibrations - Assessing item level thresholds and ordering

Table 36 reports the item response transition thresholds for the activities dimension. The EQ-5D-5L usual activities item covered the largest range of the theta sale, but was less sensitive at

the more severe points of theta. The PROMIS-29 and SF-36v2 items cover a similar range of the scale. **Figure 22** demonstrates that the EQ-5D-5L UA response levels were not operating as expected at the most severe level, as the fourth curve did not have a clear peak above the other levels. The other items were ordered as expected.

Figure 22: Threshold and information curves - Activities dimension



4.21.8. IRT item calibrations - Assessing item information

Table 36 includes the slope estimates for the activities dimension items. The EQ-5D-5L and SF-36 items had a lower slope than the PROMIS items, and generally provided a consistent level of information across the mild to moderate range of theta (as **Figure 22** demonstrates). It was also demonstrated that the PROMIS items provided a higher level of information according to the slopes and information profiles. A number of the items had information peaks, with a drop in information around the middle range of the severity response levels. This followed through to the overall information curve, where peaks in the information provided at different points of the theta scale (and aligned changes in the standard error) was observed (**Figure 23**). **Table 38** demonstrates that most information, and therefore sensitivity, occurred between 0 and 1.6 on the theta scale.

Table 38: Total test information at key points of the latent scale - Activities dimension

θ point[a]	Test information	Expected SE[b]
-2.4	1.15	0.93
-1.6	2.12	0.69
-0.8	13.91	0.27
0	22.64	0.21
0.8	21.05	0.22
1.6	23.25	0.21
2.4	8.4	0.35

[a] Selected severity points on underlying theta scale; [b] standard error

4.21.9. IRT item calibrations - Assessing IRT score estimates

Figure 24 reports the estimated score curve for the activities dimension (total score = 28). A theta of zero was equivalent to score of 7, one was equivalent to 14, and two was equivalent to 22. The overall increase in information was less steep than for the other three dimensions tested.

Figure 23: Total information curve - Activities dimension

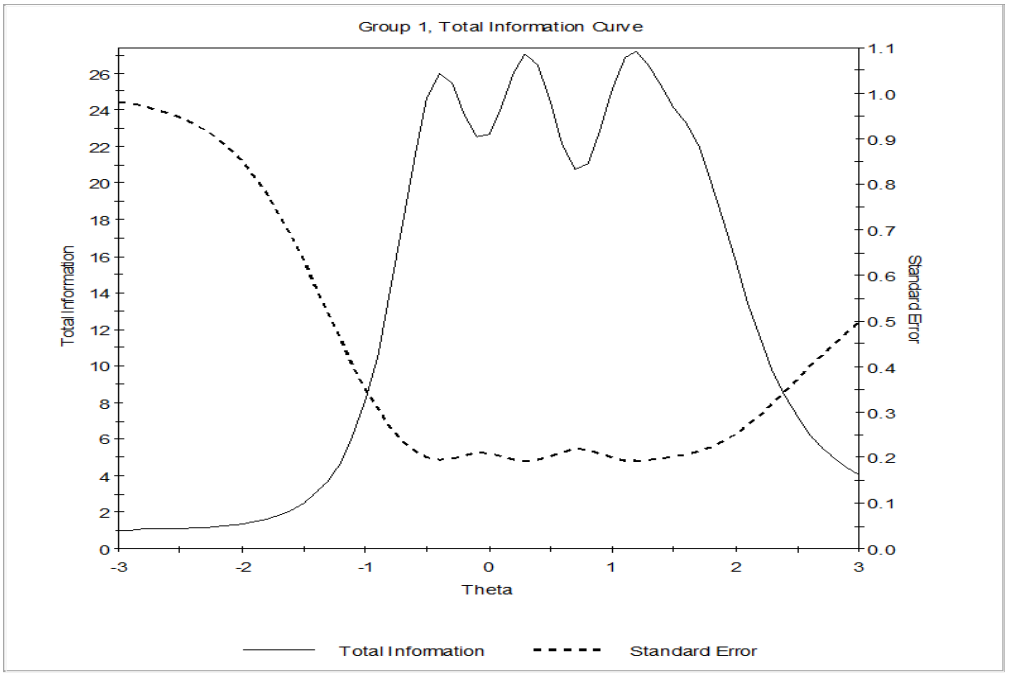
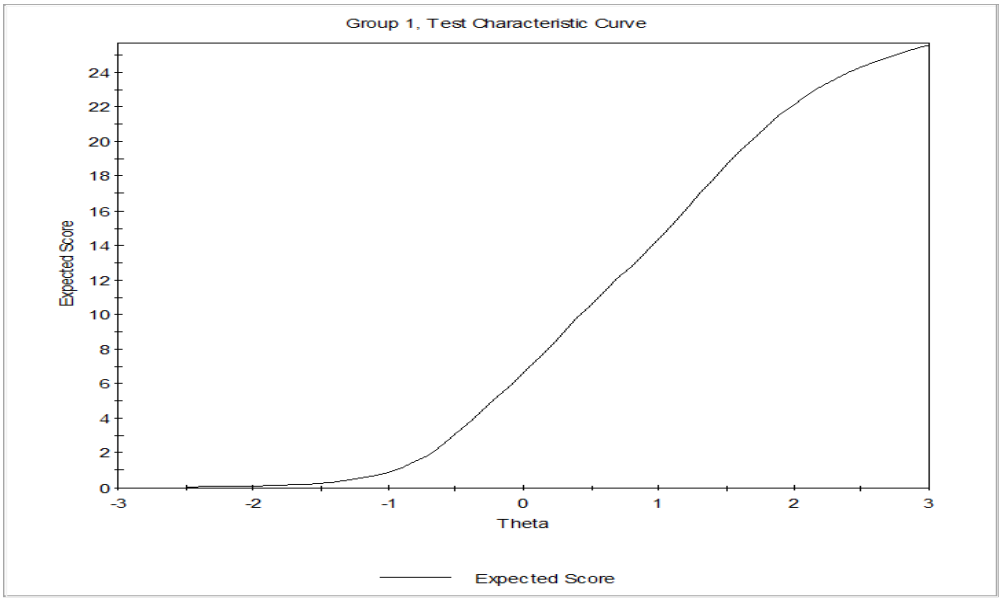


Figure 24: Estimated score curve - Activities dimension



4.21.10. Summary and implications of results

Regarding the activities dimension, there was some evidence of misfit for a selection of the items, and this could be linked to the status of these items as non-loaders in the EFA. However, the overall dimension fit was satisfactory. The EQ-5D UA item does again provide information across the central range of theta, demonstrating that it could be used as a general indicator of the level of activity impairment. Further items would inform the development of this dimension, and increase the sensitivity of the item pool to different levels of activity limitations.

4.22. Overall discussion and implications for extending the QoL measurement framework

4.22.1. Summary

The analysis conducted in this chapter used multiple methods based in IRT to understand the relationship between QoL instruments that are drawn from a broader framework than is typically used in the economic evaluation of health care. The potential for moving the measurement of QoL forward using innovative approaches that derive from a broad composite measure of QoL outcomes (Extension 1), and also combining preference-based measurement with the elicitation of more detail about each dimension included (Extension 2), was investigated. The focus was on assessing dimensionality and using IRT based analysis methods. The results provide information about and implications for the wider measurement of QoL, and for the development of a more flexible layered approach to measuring preference and non

preference-based outcomes. These issues, the limitations of the analysis, and areas for possible further work, are described below.

4.22.2. Implications for the measurement of QoL

The results of this study provide information regarding the relationship between items. There are dimensions identified that map onto the existing HRQoL dimensions, including physical functioning and mobility, mental health and pain. Physical functioning was a consistently highly loading factor including EQ-5D, SF-36v2 and PROMIS-29 items. Of note is the suggestion that three EQ-5D-5L dimensions (MO, SC and UA) load within the same dimension. This raises questions about the independence of these dimensions from a measurement and valuation perspective. Mental health and pain are also generally consistent, and are essential dimensions to assess in any measure of QoL, as evidenced by extensive qualitative and quantitative measure development work [114, 115]. Another HRQoL area of importance is social functioning and activities. The items assessing these issues did not form a clear dimension, but a pattern of non or cross loading of these items could be observed. This has been seen elsewhere [193], and may be due to the way in which the questions are asked, which include broader concepts and examples than the items included in physical functioning and mental health, which ask more directly about particular concepts. The dimensionality assessment results suggest that these items are impacted by both physical and mental health, but when calibrated within a unidimensional model there is evidence that an underlying construct of activities is being measured

A range of issues are raised by the suggested dimensionality of the wider health and QoL concepts. One consistent factor across analyses is a combined dimension including a large number of ASCOT and ICECAP items. This may not be expected given that those measures were developed to assess SCRQoL and capabilities, which may have some conceptual overlap but also differences in what is measured, and how the items are framed. However, it does demonstrate that what is assessed is different to the concepts included in HRQoL frameworks. The relationship between wider measures of QoL requires investigation in a range of samples with different conditions.

There is also a dimension assessing role functioning as a result of both physical and mental health problems. Items from this dimension are included in the SF-6D, which seems to build on the measurement of usual activities as described by the EQ-5D-5L by asking about activity

limitations in a different way. The EQ-5D-5L provides examples of usual activities as part of the overall item, and we do not know what aspect of usual activities the respondent is responding to. In contrast the SF-36 asks about different activity and functioning concepts directly across different items with associated response levels. The approaches provide complementary data for the assessment of activity and functioning impairments.

Another dimension that is identified that is part of the SF-6D is energy/vitality which includes items that are positively worded. The counterpoint to this is the identification of a negatively worded fatigue dimension. This modelling of factors covering similar concepts, but clustered in terms of the direction of the wording, has been found elsewhere [198]. It demonstrates the importance of wording in the generation of items, but also a possible limitation with the dimensionality assessment approach, where items with the same wording direction, but also items using the same response patterns, may cluster together. If a wider measure was to be developed from this dimensionality, only one of these overlapping dimensions would be required.

A further dimension includes the four PROMIS items focused on sleep. Neither the EQ-5D-5L or SF-36 include items around sleep, which may be an area for development within a wider measure of preferences [196].

A general wellbeing dimension includes a number of items from the WEMWBS. It may be that the wellbeing items do cluster together in terms of what they are measuring, and this raises questions about how wellbeing could be included in a broader measurement framework.

There is also a number of less clear dimensions that, are more challenging to define. The first of these is around self-care and related issues. This includes a question from the EQ-5D-5L, but the limited items in the wider dimension, and the divergence of the issues covered by the questions, means that IRT analyses of the item pool in this dimension would not be informative. A second unclear dimension is around dignity and items assessing wider concepts including clarity of thinking. These issues may be more important for certain population groups.

4.22.3. Implications for developing a flexible approach to measuring outcomes

The broader dimensionality identified by the Extension 1 analyses also provides a basis for testing how a more flexible approach to measuring preference and non preference-based

outcomes could be conceptualised and operationalised. This could be tested using HRQoL dimensions that are already widely included in existing measures. In the analysis conducted here, four commonly assessed dimensions that included EQ-5D-5L items along with items from other measures assessing the same constructs were tested. The aim was to explore the items and dimension relationship, and the feasibility of the approach, rather than refine the item sets to produce a usable measure. Therefore, no items were removed at this stage, but instead the results were used to understand the feasibility of the development process, and indications of how layered dimensions might appear based on existing items.

However, this process also produced a calibrated item bank with theta scores which further demonstrates how a layered approach could be scored and operationalised. It also demonstrates how a layered approach could be implemented based on the theta scores that are assigned to sets of raw scores within dimensions using a CAT based approach. In this operationalisation, the first layer provides the utility score, and the second layer, which provides more detailed dimension-relevant information, could be implemented using CAT methods. This was not tested here as the aim was to examine the feasibility of the approach. In each dimension, the completion of the EQ-5D question could be used to generate a utility, and the raw item and theta scores from completing each item bank would provide more information, with the standardised theta score comparable across dimensions, and scalable based on condition severity to inform wider health decision making. This framework could also benefit from the wording of the EQ-5D-5L items, which are quite general (for example asking about anxiety and depression and usual activities), in comparison to the items on other measures which often include multiple items assessing concepts using more specific positive and negative wording (so therefore are amenable to a layered approach as they can provide more focused and detailed information).

4.22.4. Study limitations and suggestions for future research

This study is not without limitations, and associated areas for future research. A feature of the study that has both advantages and disadvantages is the use of items from existing instruments. This approach is advantageous as it was not possible to develop items across many areas of QoL *de novo* to allow for a test of IRT methods, and the development of a layered approach to outcome measurement. A number of past studies [85, 192, 193] have applied IRT methods to existing measures to test dimensionality, and develop PBM descriptive systems. The PROMIS item bank development was also based to a large extent on the use of existing items [208]. This

study adapts those methods to test a wider set of QoL items. However, using existing measures is also a disadvantage, as the analysis is informed by the characteristics of the existing items. The dimensionality results are also linked to measures chosen, and there are many that could be included. However, due to respondent burden it is difficult to include too many measures. Further work can repeat the analysis in other datasets including multiple measures, or develop items *de novo*.

An associated limitation is the health severity distribution of the data. This was a result of the data collection method, which was conducted online, recruiting via a panel that targeted people self-reporting a health condition plus members of the general population. The recruitment of a sample covering a narrow severity range could impact the consistency of the findings for items that are sensitive to more severe problems (e.g. self-care limitations) and diverse areas of QoL. Recruiting a wider severity distribution would strengthen the results for these items. However, the main strength of the data and sampling approach was the ability to collect a diverse range of measures from a sample that are unlikely to be collected together in other clinical settings.

The majority of the dimensions tested using IRT methods were identified by the dimensionality assessment. This was not the case for the activities dimension, where the items did not load on any dimension. However, the results of the analysis on the activities dimension did demonstrate unidimensionality. Further work could attempt to understand why activities items do not load with each other. One possible explanation for this is because the concept of activities can be represented and measured using diverse examples (given the diversity of activities that can be assessed that include, for example, work, leisure and family). In contrast consider the physical functioning dimension, where the functioning examples measured within items across measures have similarities (for example walking is included in multiple measures).

An associated criticism, and possible explanation for the dimension model characteristics, is linked to differences in the conceptual reduction of diverse QoL dimensions to a latent trait. The similarities in what is measured by the physical functioning items facilitates the reduction of these concepts to a latent trait, but this is more challenging for the broader dimensions which include more diverse descriptions, for example the combined social care QoL dimension. Given the aim of Extension 2, IRT was not tested on this dimension, so the level of validity of the method for these descriptions is an open question. However, the IRT work on the broader activities dimension (which could be criticised for not reducing to a latent trait) was feasible, and

produced meaningful results. To avoid disenfranchising certain domains of QoL, a range of psychometric methods should be combined in the development process.

In this study, item dimensions were tested, but not fully calibrated, as poorly fitting items were not removed. However, the aim of this study was to explore the relationship, and not develop and refine dimensions for wider use. Therefore, the analysis provides information on the relationship across the items, and how to extend the measurement of outcomes to increase the QoL dimensions included, and the QoL information collected within existing dimensions.

There was also evidence of local dependence that would require careful consideration, and the removal of overlapping items, if this approach was used in further measure development. The evidence of local dependence across items within the same measure highlights issues with these that could be considered in any future assessment of the validity of the instruments.

4.23. How this study informs this thesis

In this study, the measurement relationship between a range of QoL outcomes has been investigated. In the chapter that follows, this work is extended to understand the valuation relationship between HRQoL and wider QoL outcomes using DCE methods. This work will inform the feasibility of valuing wider outcomes in the same framework. It also allows for an understanding of how respondents value wider outcomes. This is followed up by an examination of design methods for DCE studies, to inform the most appropriate way to design studies for the valuation of health and wider QoL outcomes. Combining the measurement and valuation evidence allows for a wider assessment of how to broaden the instruments and values used for resource allocation decision making.

5. Testing the performance of existing preference elicitation methods to develop a value set for a measurement system combining health and social care related quality of life

5.1. Summary

In Chapter 4, the measurement relationship between instruments assessing QoL from diverse perspectives was investigated. The results provide information about how the measurement of health and QoL could be broadened. As described in Chapters 1 and 2, both the measurement and valuation of outcomes are essential parts of the development of outcome measures. Therefore, any broader measurement framework that was developed would require valuation using methods that allow for diverse domains to be valued using the same metric. Such a broader instrument could encompass the QoL dimensions found in the results of Chapter 4, but could also focus on a subset of the broader dimensions.

In this chapter, the combined valuation of a subset of broader QoL domains is investigated using DCE. The broader framework that is valued includes dimensions measuring HRQoL (using EQ-5D-5L) and SCRQoL (using ASCOT). This provides an example of the feasibility of the process of valuing a broader instrument, and has applicability for the valuation of other QoL frameworks. The results of this study will provide evidence about population preferences for broader QoL concepts. The results will help understand whether developing a broader measurement framework including a value set informed by population preferences for diverse QoL outcomes is feasible.

5.2. Introduction

As described in Section 2.8, value sets estimating HRQoL weights for use in the calculation of QALYs are often developed by eliciting the preferences of the general population for health states described by generic health-focused PBMs. A challenge that is not addressed by health-focused PBMs is the increasing recognition by consumers and decision makers that the impact of many health and care interventions extends beyond a narrow definition of health outcomes to include broader non-health and QoL impacts [119, 120,223].

In a number of population groups or people with certain conditions in aged care, palliative care, disability care and vulnerable populations more generally, this includes issues around social care, and related impacts on QoL. The interplay between health and social care is fundamental to facilitating longevity, and maintaining or improving QoL during this time. The importance of

social care means that there are settings and circumstances where the health-focused QALY is insensitive as it does not capture the full value of an intervention or care setting. HRQoL is likely to be affected by the health conditions that these populations experience, but the SCRQoL impacts of interventions and care settings will not be specifically measured, except to the extent that they are reflected in changes in HRQoL. Consequently, SCRQoL will not be reflected in the values used for decision making based on comparing the relative performance of interventions if a measure of HRQoL is the source of the utility values. The results and summaries of submissions to the PBAC regarding particular interventions demonstrates this issue further (see **Box 1**).

Box 1: Examples from PBAC where existing MAUIs did not capture relevant impacts

(1) Icatibant for hereditary angioedema –*PBAC noted that the benefits related to* ***increased security*** *and* ***control*** *from the availability of the treatment rather than the health gain from treatment of attacks.*
(2) Poly-L-Lactic Acid for facial lipoatrophy –*PBAC noted the importance of* ***social*** *and psychological impacts were not captured by the SF-6D.*
(3) Tobramycin inhalation powder for cystic fibrosis –*PBAC noted the* ***heavy burden*** *of standard treatment and the value of* ***a transportable easy to use*** *device, factors not captured in standard health outcome measures.*

Recently, instruments that focus more directly on SCRQoL issues that arise from care interventions, such as the ASCOT [126], have been developed. Studies have compared the measurement properties of the EQ-5D-3L and the ASCOT in a variety of populations. Van Leeuwen et al [224] found that the EQ-5D-3L was more strongly associated with physical limitations than ASCOT, but less strongly associated with instruments measuring aspects beyond HRQoL in frail older people. Rand et al [225] found that the ASCOT utility score was moderately correlated with the EQ-5D-3L usual activities and anxiety/depression dimensions (with moderate correlations defined as 0.3 - 0.5), but correlated at a lower level (<0.3) with mobility, self-care and pain/discomfort. This work was conducted in adults with long-term physical, sensory and mental health conditions. In a community-dwelling sample, Kaambwa et al [226] found that only the pain/discomfort dimension was moderately correlated with the ASCOT utility score (with moderate in this study defined as 0.4 - 0.6), with correlations between the EQ-5D utility score and the ASCOT dimension scores ranging from low (<0.4 for control, cleanliness, occupation and accommodation) to moderate (for food and drink, safety, social participation and dignity). Content validity assessment suggests that respondents considered the items of both instruments valuable, but neither provided a comprehensive picture of a patient's QoL

[227]. This evidence provides some support for the use of the EQ-5D and ASCOT as complementary outcome measures for the economic evaluation of relevant health and social care interventions in different populations.

As yet, little work has investigated the relationship between instruments measuring diverse areas of QoL when both are valued using the same preference elicitation method on the same scale. This issue is investigated in the study reported in this chapter. The basis of this work is the proposition that unifying the constructs included across instruments such as the EQ-5D-5L and ASCOT could lead to the development of measures and methods that allow decision makers to assess value for money in an inclusive and consistent way across a wider range of disease and treatment contexts. Doing this would facilitate the assessment of both health and social aspects of QoL on the same utility scale. If feasible, comparisons between interventions that have impacts predominantly on HRQoL, on SCRQoL, or on a combination of both, would be facilitated. Combining instruments could extend the assessment of QoL impacts to cover wider QoL issues such as those that were suggested as missing in the PBAC submission summaries described in Box 1. This could lead to improved confidence in the utility values available for resource allocation decision making.

The theoretical utility proposed for development and testing in this study is a latent (unanchored) utility that combines preferences for health (HRQoL) and social care (SCRQoL) aspects of living. Leading from the full health to dead HRQoL utility scale discussed in Chapter 2, the best state described in this framework would be equivalent to having no problems with HRQoL and the ideal SCRQoL situation (as measured the instruments from which the dimensions included in the overall classification system are taken). The worst state on the combined latent scale describes the most severe HRQoL state, and the worst social care situation. The theoretical utilities estimated are not anchored on the QALY scale, but demonstrate the relative importance of domains of HRQoL and SCRQoL that could inform such a scale. They could be used to understand the relative importance of different QoL concepts within the combined framework, and inform preferences for aspects of interventions with health and social care impacts.

The derivation of preferences capturing different aspects and benefits of HRQoL and SCRQoL simultaneously using the DCE valuation method (that as Chapter 3 demonstrates has been used widely in the development of value sets) has not been tested. This chapter reports an exploratory DCE study, which collected preference data from an Australian community sample,

to investigate the joint valuation of HRQoL, as measured by the EQ-5D-5L, and SCRQoL, as measured by ASCOT. The EQ-5D-5L was chosen to represent HRQoL as it is widely used, and provides simple and consistent descriptions of each of the five HRQoL dimensions (see Section 2.11.3). The ASCOT was chosen as it is the predominant PBM measuring factors related to SCRQoL.

This study makes an important contribution to the emerging literature exploring approaches to valuing interventions that go beyond the HRQoL focus of the QALY [120]. Investigating relative preferences across the EQ-5D-5L and ASCOT also provides evidence about the use of DCE for the potential development of a combined instrument with a value set informed by health and wider non-health aspects. This could be extended beyond SCRQoL, but SCRQoL is suggested as a useful concept to test this with given it is important in a range of populations. Furthermore, the results in Chapter 4 suggest that there is divergence in the measurement framework of each instrument. The results may have wider applicability to other emerging work in this area (such as the development of the 'extended' QALY measure incorporating a range of QoL aspects [31]).

5.3. Aims and objectives

The study reported in this chapter links to Aim 4 of this thesis which is to test the use of DCEs to develop a value set for a combined measurement system assessing different concepts of health and QoL. This includes the following two objectives:

1. To test the use of DCEs to elicit preferences for QoL profiles that incorporate aspects of both HRQoL and SCRQoL as measured by the EQ-5D-5L and ASCOT.
2. To investigate the relative magnitude of preferences for different aspects of HRQoL and SCRQoL as measured by the EQ-5D-5L and ASCOT.

5.4. Methods

In this study, a DCE which combined the EQ-5D-5L and the ASCOT was developed and implemented in an Australian general population sample. The development, construction and administration of the DCE, and the subsequent analysis of the data, are described in detail below.

5.4.1. Development of the DCE valuation task

The DCE format used in this study was developed based on the dimensions of the EQ-5D-5L and ASCOT. The wording of the dimensions as they are presented in the original measures is

displayed in **Table 2** and **Table 4**. The DCE choice sets were developed to present pairs of QoL states comprising dimensions from both the EQ-5D-5L and ASCOT (13 dimensions in total) and asked respondents to choose which profile they preferred. A pairs format was used as it has been demonstrated to be amenable to completion by respondents, and it is a widely used format for the estimation of values. It can also be analysed using a range of models with different assumptions about respondent preferences, and this enables the relationship between the QoL dimensions to be demonstrated in different ways. It was also decided to develop the DCE tasks without including duration, either fixed across pairs to act as a frame of reference for respondents, or differing between profiles to allow for anchoring on the utility scale. This decision was made to focus trade-offs on the QoL descriptions rather than incorporating duration (which has been found to be a key driver of choices when developing value sets) [55] into the decision making process. This produced latent scale DCE values. The inclusion of duration can anchor values onto the utility scale. Below the two QoL descriptions, respondents were asked which of the two health states they preferred (given that the focus was on preferences for different QoL states).

The profiles presented included all five EQ-5D-5L dimensions. ASCOT includes nine dimensions, but only eight of these are used to generate the value set [126]. This is because two of the dimensions ask about dignity, and only one of these was included in the valuation. For consistency with the original valuation, the same eight dimensions were included in the DCE. This means that each choice set profile included 13 dimensions. As this number is a reasonably high number of dimensions to include in a DCE profile (see Section 3.3.4), the choice sets were simplified by imposing overlap in the design [44, 152]. That is, five of the dimensions in each choice set were constrained to have the same level of severity across both profiles. The levels of the other eight dimensions varied in each task. To make this clear, the dimensions that differed within choice sets were highlighted with a light grey background. An example choice set can be seen in **Figure 25**. The use of shading or colouring to highlight dimension level differences within choice sets has been shown to produce similar choice results to those obtained without shading, whilst simplifying the choice set for respondents [43]. The information and instructions provided to respondents prior to presenting the choice sets was developed based on previous DCE studies, and iteratively refined by the author and supervisory team.

Figure 25: Example DCE choice set

Please consider and imagine living with the two health descriptions below. Then tell us which description you would prefer to live in.

Health description A	Health description B
You have no control over your daily life	You have no control over your daily life
You feel adequately clean and presentable	You feel clean and are able to present yourself the way you like
You don't always get adequate or timely food and drink	You don't always get adequate or timely food and drink
You feel as safe as you want	You feel as safe as you want
You have some social contact with people, but not enough	You have little social contact with people and feel socially isolated
You are able to spend time as you want, doing things you value or enjoy	You are able to do enough of the things you value or enjoy with your time
Your home is adequately clean and comfortable	Your home is adequately clean and comfortable
The way you are helped and treated completely undermines the way you think and feel about yourself	The way you are helped and treated does not affect the way you think or feel about yourself
You are unable to walk about	You have moderate problems in walking about
You have slight problems washing and dressing yourself	You have severe problems washing and dressing yourself
You have moderate problems doing your usual activities	You have slight problems doing your usual activities
You have no pain or discomfort	You have extreme pain or discomfort
You are moderately anxious or depressed	You are moderately anxious or depressed
Which do you prefer?	
Health description A	Health description B

5.4.2. Pilot study to test survey functioning

Although the results of past research into the effectiveness of dimension level overlap [44,228] and formatting [43, 44] were used to support the choice set development, it was still possible that the format was not feasible to respondents. This could have resulted in poor quality choice data, and unexpected model patterns. An initial pilot launch phase was conducted to assess the functioning of the survey and the feasibility of the choice set format to respondents. This was measured via multiple choice usability questions about the difficulty of the tasks (including the overall difficulty, the difficulty imagining the descriptions, and the difficulty telling the difference between them) alongside a free-text question to understand respondents' opinions of the survey questions and the content in general. Initial modelling of the DCE data was also conducted, where indications of coefficient ordering between the levels of each dimension were assessed. The completion times for the pilot sample were used to inform the minimum survey completion time imposed for the full sample.

5.4.3. Study design – Constructing the design

As per previous studies employing DCE methods to value QoL instruments [48, 57], the design that was developed was specified to include substantially more choice sets than there were

parameters in the model to estimate. The main effects model combining the EQ-5D-5L and ASCOT has 44 parameters calculated as follows:

EQ-5D-5L: (5 dimensions x (5-1 levels as baseline parameter is not estimated)) = 20 +
ASCOT (8 dimensions x (4-1 levels)) = 24.

The overall design developed included 300 choice sets which is greater than six times the number of parameters required for the standard main effects multinomial logit (MNL) model, and more than four times the size of any other model reported in the results. The design was constructed using a modified Fedorov algorithm. The objective function of the algorithm was to optimise the estimation of the main effects model using the criterion of minimal D-Error. The algorithm iteratively improves the set of choice sets included in the design, with improvement measured by reductions in the D-Error [229]. The design was implemented using the DCE design software Ngene [67], which was set to iterate through designs and improve sequentially for 24 hours.

To allow for the design to include overlap, a large candidate set of choice sets with overlap on five dimensions was used as the basis for the design. This is a requirement of selecting an overlapping design when Ngene is used. The candidate set included 300,000 choice sets with overlap on five dimensions, but unrestricted on which dimensions and levels were overlapping. This was developed based on the full factorial of the design listing all possible combinations. The candidate set was linked to Ngene, and the Fedorov algorithm [241] was applied to the candidate pool to select 300. The Fedorov algorithm randomly selected a design of 300 choice sets, and iteratively changed profiles with others in the candidate in sequence, retaining any improvements based on minimising the D-Error of the overall design. No guidance exists as to the required relative size of a candidate set in comparison to the number of choice sets to be included in the final design. In this study, 0.001% of the candidate sets were included in the final design.

5.4.4. Study design – Use of zero priors

Prior information, if available, is often used to inform the design process. In this study, the design was not informed by priors given that this is the first study to attempt to value both EQ-5D and ASCOT in the same DCE framework. This is described as using non-informative, or zero, priors, and it has been suggested that zero priors are particularly useful if valid or known priors are not

available (as in this study) [231]. Priors could have been taken from studies that valuing the instruments separately using DCE, but this would not have taken into account the potential impact on values of the trade-off between dimensions which is a key concept that is tested in this study.

5.4.5. Study design – Issues around implausibility

There is the potential for implausible combinations of health and social care dimension levels, for example having no problems with self-care and not feeling at all 'clean or presentable'. However, no combinations of HRQoL and SCRQoL were restricted in this study. This is because it is difficult to make a judgement, a-priori, that certain combinations are not realistic, particularly as what is considered implausible has been found to be respondent specific [232]. Marten et al [233] found that EQ-5D-5L level combinations assumed to be implausible (for example no problems with self-care combined with unable to do usual activities) actually appear in general population self-report data. Excluding particular level combinations may also lead to imbalance in the design (where certain combinations of dimension levels do not appear as frequency as others) and impact the coefficient estimates derived from subsequent model estimation in non-systematic ways [50]. Furthermore, implausibility in relation to combinations of the ASCOT dimensions has not been investigated and combining different areas of QoL makes researcher judgements about which level combinations are implausible more complex.

5.4.6. Study design – Blocking and dimension ordering

The 300 choice sets were separated into 20 blocks of 15 using the blocking functionality available in Ngene. This blocking function allocates the choice sets to blocks to balance dimension level occurrence within the block. Each of the 20 blocks was included in two versions of the survey that replicated the full design: Version 1 presented the EQ-5D-5L dimensions followed by the ASCOT dimensions and Version 2 presented the ASCOT dimensions followed by the EQ-5D-5L dimensions, with the dimensions within instruments presented in the standard order described in the introduction. Respondents were subsequently randomised to one of 40 survey blocks. As an example, Appendix 11 includes one of the blocks of choice sets from the design. Further randomisation of dimensions within the DCE could have been imposed, but the decision was made not to do this to allow respondents to always see concepts related to HRQoL (EQ-5D-5L) followed by SCRQoL (ASCOT) or vice versa. Evidence for dimension order effects in previous health state valuation work is inconclusive [148].

5.4.7. Survey design and administration

The survey was administered online, which is a widely used approach to the collection of DCE data in Australia and internationally. The survey comprised background information about the project and ethics approval, followed by an informed consent page, then questions on respondents' demographic characteristics, health and QoL (including EQ-5D-5L and ASCOT to collect data on the respondent's QoL, but also to introduce them to the dimensions they would see in the DCE tasks). Subsequently, respondents were shown instructions about the task (see Appendix 12) and were told that they will see two different descriptions of health and social care, a warm up task followed by the 15 DCE tasks. The usability and free-text pilot questions were also included in the main study. The order of appearance of each set of tasks within a Block was randomised. The full survey content is included in Appendix 12.

5.4.8. Recruitment and respondents

The study aimed to recruit 1,000 respondents from the Australian general population, targeted to be representative in terms of age (across six categories defined as 18 – 24; 25 – 34; 35 – 44; 45 – 54; 55 – 64; 65+) and gender. A representative sample was sought to mimic the samples that are used for value set development. Including respondents from different demographic groups also allows for an understanding of preferences for diverse QoL concepts across different population groups. An overall sample of 1,000 respondents was targeted to provide approximately 50 observations per DCE choice set (1,000 respondents x 15 observations per person divided by 300 choice sets overall). This number of observations per choice set is in line with other DCE studies developing value sets (see Chapter 3). Respondents were required to complete the survey in longer than a minimum completion time of 3 minutes.

The initial pilot launch aimed to recruit approximately 10% of the overall sample. The survey was then reopened following initial assessment of the survey functioning and responses to the DCE choice sets. Respondents were recruited at random from existing internet panels managed by Survey Sampling International, who allocated respondents who were willing to complete questionnaires during the data collection period. The panel company recruited from multiple subpanels with different respondent demographics, but under their broad management, to support the generalizability of findings.

After entering the survey, respondents read the project information and provided implied consent. The questions were then started. A small incentive was provided if respondents

completed the full requirements of the survey in more than the minimum completion time. The respondents were not informed about the use of a minimum completion time. The amount and type of incentive differed depending on the procedures of the subpanel from which the respondent was recruited. This methodology was approved by the Centre for Health Economics Research and Evaluation Program Ethics Process for low risk projects (UTS HREC REF NO. 2015000135).

5.4.9. Data analysis and modelling – Sample

The demographic characteristics of the sample were compared to those observed in the Australian population. The EQ-5D-5L and ASCOT utility scores were also calculated and assessed to provide a measure of distribution of HRQoL and SCRQoL within the sample. For the EQ-5D-5L the Australian value set developed by Norman et al [57] was used (see Section 2.11.3). For ASCOT the UK value set [126] was used as an Australian value set is not available (see Section 2.13.2). The frequency of respondents answering each severity level of each of the EQ-5D-5L and ASCOT dimensions was assessed to understand how the general population respond to the dimensions.

5.4.10. Data analysis and modelling – Conditional logit

Initially, the data were analysed using conditional logit regression which generated coefficient estimates for each level of the EQ-5D-5L and ASCOT dimensions. For a full explanation of the conditional logit model see Section 2.9.10. Robust standard errors were used in the model to take into account that each respondent provides multiple observations. Conditional logit regression allows for comparison of the overall magnitude (a proxy for importance at the overall level) of the dimensions included. This allowed for the overall rank of the 13 dimensions to be assessed. The data were modelled for the whole sample (**Model 15**), and a consistent version of this model that imposed ordering on the levels within dimensions, were estimated (**Model 16**). The consistent version was generated as an example of a model that would be suggested for the calculation of utilities, with monotonic estimates forced across dimensions.

Conditional logit regression was the natural starting point for analysis, as coefficients can be interpreted as decrements away from the baseline level to give an indication of preferences across the dimensions and levels for the overall sample. This is possible given the assumption that the overall sample has the same underlying (homogeneous) preferences. However, the assumption of preference homogeneity can be criticised as it is unlikely to be true for the

dimensions included in the choice set developed in this study. For example, how likely it is that preferences for avoiding certain dimensions of health and social care are same across respondents in different age groups, with different experiences of ill health, and different social and family situations? To counteract this issue, the analysis was extended to include models that take heterogeneity of preferences into account. However, before that a series of interactions between HRQoL and SCRQoL dimension levels were explored

5.4.11. Data analysis and modelling – Testing interactions

To understand preferences for the relationship between the HRQoL and SCRQoL dimensions in more detail, interactions between sets of attributes from each instrument were developed, and incorporated into the model alongside the main effects parameters using conditional logit. The interactions were developed to estimate how having a high level of problems on HRQoL and SCRQoL were valued when there were no problems on the dimensions of the other concept. This tests whether having a range of SCRQoL concerns is perceived as more manageable when HRQoL is not problematic, and vice versa. To test this for SCRQoL, dummy variables interacted a combination of all appearances of level 4 ASCOT dimensions with all appearances of each of the EQ-5D dimensions when they had no problems, so producing five interactions described as: N1_MO x N4_ASCOT; N1_SC x N4_ASCOT; N1_UA x N4_ASCOT; N1_PD x N4_ASCOT; N1_AD x N4_ASCOT. To allow for a comparison of the magnitude of the coefficients, an interaction indicating any appearance of a level 4 ASCOT dimension was also estimated. This is similar to interaction terms used in other value set development modelling [57], and allows for an assessment of preferences for severe levels of SCRQoL at an overall level.

The matched interactions for severe HRQoL included dummy variables combining all appearances of Levels 4 and 5 for the EQ-5D-5L interacted with level 1 appearances for each of the ASCOT dimensions, so producing eight interactions overall (N1_CO1 x N4/5_EQ5D; N1_CL x N4/5_EQ5D; N1_FD x N4/5_EQ5D; N1_SA x N4/5_EQ5D; N1_SP x N4/5_EQ5D; N1_OC x N4/5_EQ5D; N1_AC x N4/5_EQ5D; N1_DIG x N4/5_EQ5D). A comparator N45 term interaction was included when any EQ-5D-5L dimension was at level 4 or 5.

The interaction analysis was exploratory as the design was not specified within Ngene to estimate the extra terms included. However, even without full specification within the design, the models can be used as an indicator of the impact of the interactions in terms of the

relationship between HRQoL and SCRQoL. They could also inform future work in this area testing the relationship between different QoL concepts in more detail.

5.4.12. Data analysis and modelling – Investigating scale differences between subsamples

DCE values such as those generated by conditional logit models are estimated on a latent scale. Of interest in the analysis of DCE data are potential differences between the values of subsamples within the data in terms of the underlying scale of the models. For example, this could be based on demographics, or features of the study design. However, given values are on a latent scale, it is difficult to directly compare the magnitude of the values across subsamples. This issue can be resolved using analysis that assesses whether the underlying scale of responses (and therefore respondent preferences) differ between subsamples.

In this study, a scale differences were tested across a range of subsamples. Firstly, the impact of varying the order in which the measures were presented within the DCE was tested. This established whether presenting EQ-5D-5L or ASCOT first within the choice set impacted the preference estimates. Secondly, scale differences based on a number of demographics including age, gender and condition were tested. These were chosen as it was hypothesised that they may lead to different patterns of preferences across the diverse dimension descriptions included in the DCE. They were also used in other heterogeneity analyses reported in this thesis (see latent class analysis which is described in section 5.4.15). The scale model was also used to test differences based on time taken (see section 5.4.13).

The analysis used adapted the scale testing approach proposed by Swait and Louviere [71]. This approach is described in detail in Section 2.9.11 and uses an LR test to examine the null hypothesis that the underlying scale did not differ across subsamples. Stata 15 [66] was used for this modelling, with the scale model estimated using *clogithet*, a user written Stata module [174, 175].

5.4.13. Data analysis and modelling - Time taken

A key issue with the use of online methods to administer DCE studies is the lack of researcher control over the attention a respondent pays to the completion of the survey. This could result in poor quality data, and therefore impact on the validity of the preference estimates elicited. The level of respondent attention can be proxied via the assessment of the time taken to complete each choice set (i.e. removing individual tasks) and the overall survey. Both quick and

slow completion times could be indicative of a lack of attention. In this study, analysis of time taken at both the choice set (removing individual responses) and overall (removing respondents) level was conducted as a sensitivity analysis. This was done using conditional logit and scale testing models to examine consistency, and susceptibility to differences based on time taken.

At the task level, four different subsets of choice set responses were removed based on different completion times, and the choice sets remaining were modelled. Task subset one removed tasks completed very quickly (defined as three seconds or less), or very slowly (defined as 1,800 seconds (30 minutes) or more). Task subset two removed the fastest approximate 5% and the slowest approximate 5% of choice sets. Task subsets three and four were linked. Subset three excluded the fastest approximate 25% and slowest approximate 25%, and subset four included the tasks excluded for set 3. This allowed a comparison of consistency based on approximately 50% of the sample with different time completion profiles, one of which (subset three) could be perceived as valid completers, in comparison to subset four, where fast and slow completions could be perceived as problematic (and might lead to more inconsistent models).

At the overall completion time level, three subsets of respondents were removed. Respondent subset one removed those completing in the fastest 5% and slowest 5%. Respondent subsets two and three were linked, where set two excluded the fastest 25% and slowest 25%, and set three excluded those completing between 25% and 75% overall.

Alongside this analysis, further modelling using scale testing was conducted. At the task level, the completions were split into two groups based on the median completion time (generating subgroups of task completions including approximately 50% of tasks), and scale differences between the fastest 50% and the slowest 50% were investigated. At the overall respondent level, the sample was split into two based on the median time taken generating subsamples of overall completions including approximately 50% of the respondents. The median was used to divide the sample over other measures of central tendency such as the mean, as the distribution of time taken would result in different proportions of respondents appearing in each subsample, and therefore potentially result in an invalid comparison.

5.4.14. Data analysis and modelling – Preference Heterogeneity

Exploration of preference heterogeneity was considered to be important given the range of dimensions included, which may have different impacts and meaning, and therefore preference

patterns, in different population groups. For example, age and experience of health conditions may be a factor in preferences towards certain aspects of both HRQoL and SCRQoL. To assess preference heterogeneity, both latent class [234] and mixed logit regression models [73] were used.

5.4.15. Data analysis and modelling – Latent class

Latent class analysis is used to look for groups of respondents with similar patterns of preferences. For a full explanation of the implementation of the latent class model, see Section 2.9.11. In this analysis, models were produced including parameter estimates for different preference class structures. These were estimated from the overall dataset. Models including between two and six classes were produced, with the class structure assessed for comprehensibility and preference patterns. In line with Train (2008), the optimum number of classes was determined by assessing the AIC, BIC, and the Consistent AIC.

It is also possible to extend the analysis to include parameters indicating the class membership of different demographic groups (estimated as class delimiters), and this was done here for all models. The demographic groups were entered as binary dummy variables to allow for interpretation of the probability estimates, and included age (18 – 65 years old, and 65 or older), gender, and having a long-term health condition. These were used as key demographics where preferences for HRQoL and SCRQoL might differ, and were matched with the demographics used for the scale testing analysis reported in section 5.4.12. The user written Stata package 'lclogit' [235] was used for this analysis.

5.4.16. Data analysis and modelling – Mixed logit

As described in Section 2.9.13, mixed logit is a model that allows parameters to be specified as random (i.e. to be specified as heterogeneous), alongside the specification of fixed (homogeneous) parameters. Mixed Logit was used in this study to iteratively test both different sets of EQ-5D-5L and ASCOT parameters for heterogeneity, and also different model specifications to ensure that the estimates produced were reliable indicators of respondent preferences. Mixed logit modelling was conducted in Stata using the 'mixlogit' command [236]. **Table 39** reports the range of mixed logit models and specifications included in this chapter, and the justification for their inclusion. This leads to the main models reported in this chapter, with the other models reported in appendices.

The different model specifications used iteratively tested different numbers of Halton Draws, burn rates, parameter distributions, correlations between parameters, and model maximisation procedures. Halton Draws are sequences used in the maximum simulated likelihood method that provide increased accuracy in comparison to random draws as they use sequences that are distributed more evenly. Increasing the number of Halton Draws usually results in more accurate estimation of the variance of the estimated parameters, and in this study, the estimates from two different numbers of draws were tested. These were 50 (the default used in Stata) and 1,000. In previous work, Bhat [237] and Train [238] found that using 125 Halton Draws results in increased accuracy in comparison to random draws. In this study, models with large numbers of random parameters (up to 44) are tested, and increasing the number of draws allows for increased confidence in the estimation of complex models [239]. The burn rate is used to specify how many sequence elements should be dropped when generating the Halton Draws, and reduces the correlation between draws. At a minimum, the number of sequences dropped should be equal or higher than the number of random parameters in the model [240]. In this study, multiple burn rates were specified for the models. This included the default burn rate of 15, and also was linked to the number of random parameters in the model. For example, the models with 44 random parameters had a specified burn rate of 44.

For each parameter specification, models were estimated using both normal and log-normal distributions. Models were also estimated specifying that the random dimension level parameter coefficients were both uncorrelated and correlated. This is important to test given the potential for preference relationships across coefficients within the dimensions included a the choice model (for example correlations between preferences for different areas of HRQoL). If the parameters are specified as correlated, the mixed logit model also estimates the covariance matrix between the random parameters. Given the complexity of the models, and evidence of non-concave regions in the distribution, the 'difficult' maximisation stepping algorithm was also tested. Across all model specifications, the Log-Likelihood, AIC and BIC were used as indicators of model performance.

The iterative approach to testing model structures allowed for an assessment of different specifications of fixed and random parameters testing HRQoL and SCRQoL both separately and combined. This was done for different sets of model criteria. As a starting point, heterogeneity was assessed at the aggregate dimension level including one parameter for each of the 13 dimensions. The most complex models specified that all 44-dimension level EQ-5D-5L and

ASCOT parameters were random. Given the complexity of these overall models, separate models just specifying that the EQ-5D-5L dimensions and ASCOT dimensions were random were also tested. For further assessment of heterogeneity, the magnitude of heterogeneity based on the overall dimension level model with increased draws was also used to specify three other exploratory models. The 20 most heterogeneous parameters (of the overall 44) were estimated as a single model (using 20 provided a selection of both EQ-5D-5L and ASCOT parameters), and this was repeated for the 10 and 5 most heterogeneous parameters to understand the consistency of the models, and the strength of heterogeneity.

Table 39: Mixed logit parameter specifications

Model no	Fixed	Random	Specifications	Justification
Model 71	None	MO, SC, UA, PD, AD, CO, CL, FD, SA, SP, OC, AC, DI	50 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	Tests heterogeneity at the aggregate dimension level to provide an overall indicator
Model 72	None	MO, SC, UA, PD, AD, CO, CL, FD, SA, SP, OC, AC, DI	1,000 Halton Draws, 20 burns, normal distribution, independent (uncorrelated) random coefficients, difficult	Same specification as previous model, but increases the number of draws and burns to improve estimation performance
Model 73	None	MO, SC, UA, PD, AD, CO, CL, FD, SA, SP, OC, AC, DI	1,000 Halton Draws, 20 burns, log-normal distribution, independent (uncorrelated) random coefficients, difficult	Same specification as previous model, apart from specifying log-normal distribution of parameters
Model 74	None	MO, SC, UA, PD, AD, CO, CL, FD, SA, SP, OC, AC, DI	1,000 Halton Draws, 20 burns, log-normal distribution, dependent (correlated) random coefficients, difficult	Specifying correlated coefficients
Model 75	None	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5, CO2-CO4, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	50 Halton Draws, 16 burns, normal distribution, independent (uncorrelated) random coefficients	All HRQoL and SCRQoL dimensions specified as random to understand heterogeneity at the overall level and compare to the models specifying HRQoL and SCRQoL as random separately
M 23	None	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5, CO2-CO4, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	1000 Halton Draws, 44 burns, normal distribution, independent (uncorrelated) random coefficients	Same specification as previous model, but increases the number of draws and burns to improve estimation performance

Model 76	CO2-CO4, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5	50 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	Tests the HRQoL dimension levels for heterogeneity whilst keeping the SCRQoL dimension levels fixed
M 24	CO2-CO4, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5	1,000 Halton Draws, 20 burns, normal distribution, independent (uncorrelated) random coefficients, difficult	Same specification as previous model, but increases the number of draws and burns to improve estimation performance
Model 77	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5	CO2-CO4, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	50 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	Tests the SCRQoL dimension levels for heterogeneity whilst keeping the HRQoL dimension levels fixed
M 25	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5	CO2-CO4, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	1,000 Halton Draws, 20 burns, normal distribution, independent (uncorrelated) random coefficients, difficult	Same specification as previous model, but increases the number of draws and burns to improve estimation performance
Model 78	MO2-MO4, SC2-SC4, UA2-UA4, PD2-PD4, AD2-AD4, CO2-CO3, CL2-CL3, FOOD2-FD3, SA2-SA3, SP2-SP3, OC2-OC3, AC2-AC3, DI2-DI3	MO5, SC5, UA5, PD5, AD5, CO4, CL4, FD4, SA4, SP4, OC4, AC4, DI4	1000 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	Testing heterogeneity of most severe levels of each dimension only (as past work suggests this is where preferences differ)
Model 79	MO2, MO3, SC2, UA2-UA5, PD3, CO2, CO3, CL2, CL4, FD2, FD3, SA2, SA3, SA4, SP3, OC2, AC2, AC3, AC4, DI2, DI3	MO4, MO5, SC3, SC4, SC5, PD2, PD4, PD5, AD2, AD3, AD4, AD5, CO4, CL3, FD4, SP2, SP4, OC3, OC4, DI4	1000 Halton Draws, 20 burns, normal distribution, independent (uncorrelated) random coefficients	Top 20 most heterogeneous dimension levels from overall **M *23***. This can be compared to previous model where most severe levels suggested as heterogeneous based on previous work.
Model 80	MO2-MO, SC2-SC3, UA2-UA5, PD2-PD3, AD2-AD3, CO2-CO3,	MO5, SC4, SC5, PD4, PD5, AD4, AD5, CO4, FD4, OC4	1000 Halton Draws, 20 burns, normal distribution, independent (uncorrelated) random coefficients	Top 10 most heterogeneous dimension levels from the overall **M 23**

	CL2-CL4, FD2-FD3, SA2-SA4, SP2-SP4, OC2-OC3, AC2-AC4, DI2-DI4			
Model 81	MO2-MO4, SC2-SC4, UA2-UA5, PD2-PD5, AD2-AD3, CO2-CO3, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	MO5 SC5 AD4 AD5 CO4	1000 Halton Draws, 20 burns, normal distribution, independent (uncorrelated) random coefficients	Top 5 most heterogeneous dimension levels from the overall **M *23***

MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

5.4.17. Data analysis and modelling – Generalised Multinomial Logit Model

A range of models were tested within the GMNL framework, which allows for the incorporation of both preference and scale heterogeneity in the same model (see Section 2.9.14 for a detailed description). GMNL allows for random coefficients that are correlated and non- correlated. The model produces a Tau statistic which is significant if scale heterogeneity is present. The analysis was conducted using the user written Stata module 'GMNL' [241].

A range of models were tested, and the performance statistics and interpretability of the coefficients was assessed. **Table 40** describes the GMNL models tested, and the justification for their inclusion. The model specifications around the number of draws, burn rate, distribution, and parameter correlations were based on the results of the iterative mixed logit modelling process. The dimension level models tested were **Model 26** which allowed the 20 EQ-5D-5L parameters to vary to understand scale and heterogeneity across the HRQoL attributes. **Model 28** allowed the higher severity levels of each instrument (levels 4 (severe) and 5 (unable to/extreme)) of EQ-5D-5L, and level 4 (various descriptors from the ASCOT) to vary to test the hypothesis that preference heterogeneity is more prevalent as severity increases. **Model 29** takes the 20 parameters demonstrating the highest level of heterogeneity from the most complex mixed logit model that allowed all EQ-5D-5L and ASCOT parameters to vary (**Model 23**). This method was used to select the dimensions to specify as random, as it provides insight into the complex relationship between HRQoL and SCRQoL in terms of which dimensions and levels

exhibit evidence of heterogeneity without specifying that scale also varies within the models. In contrast to the mixed logit analysis, a model allowing all parameters to vary was not tested under the GMNL framework as the maximum number of parameters that can be specified to vary within Stata is 20. However, given the complexity of the GMNL model, 10 was seen as sufficient to test scale and parameter heterogeneity, and compare back to the mixed logit models. An issue with estimating the GMNL model is having sufficient data. There is no clear guidance regarding sample size, so the analysis reported in this chapter is exploratory.

Table 40: GMNL model specifications

Model no	Fixed	Random	Specifications	Justification
Model 26	CO2-CO4, CL2-CL4, FD2-FD4, SA2-SA4, SP2-SP4, OC2-OC4, AC2-AC4, DI2-DI4	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5	1000 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	Test level of preference and scale heterogeneity of the EQ-5D-5L dimension levels
Model 27	MO2-MO5, SC2-SC5, UA2-UA5, PD2-PD5, AD2-AD5, CO2, CL2, CL4, FD2, SA3, SP3, OC2, AC2, AC3, DI2, DI3	CO3, CO4, CL3, FD3, FD4, SA2, SA4, SP2, SP4, OC3, OC4, AC4, DI4	1000 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	13 significantly heterogeneous ASCOT parameters from overall mixed logit **M *23***
Model 28	MO2-MO3, SC2-SC3, UA2-UA3, PD2-PD3, AD2-AD3, CO2-CO3, CL2-CL3, FD2-FD3, SA2-SA3, SP2-SP3, OC2-OC3, AC2-AC3, DI2-DI3	MO4, MO5, SC4, SC5, UA4, UA5, PD4, PD5, AD4, AD5, CO4, CL4, FD4, SA4, SP4, OC4, AC4, DI4	1000 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	To test issues related to potentially increased levels of heterogeneity at more severe dimension levels
Model 29	MO2, MO3, SC2, UA2, UA3, UA4, UA5, PD3, CO2, CO3, CL2, CL4, FD2, FD3, SAFE2, SA3, SA4, SP3, OC2, AC2, AC3, AC4, DI2, DI3	MO4, MO5, SC3, SC4, SC5, PD2, PD4, PD5, AD2, AD3, AD4, AD5, CO4, CL3, FD4, SP2, SP4, OC3, OC4, DI4	1000 Halton Draws, 15 burns, normal distribution, independent (uncorrelated) random coefficients	Top 20 most heterogeneous dimension levels from overall **M *23***

MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

5.4.18. Assessing model performance

A number of model performance indicators were used to assess the models estimated in this chapter, namely the AIC [221], and the BIC [222]). The AIC is based on in-sample fit and estimates the likelihood of a model for estimating the specified values. Generally, a model with a lower AIC is preferred, although it is also important to assess the interpretability of the coefficient estimates produced. AIC is calculated as in Equation 12, where *K* is equal to the number of parameters in the model, and $\hat{L}$ is the likelihood.

$$AIC = -2ln(\hat{L}) + 2K \quad (12)$$

The Log-Likelihood is a commonly used measure of model fit used across all the models produced in this chapter. The Log-Likelihood is also reported across the models, but this is sensitive to sample size so cannot be used as a direct measure of fit. In general, higher values are better given that the models aim to maximise the likelihood. The BIC is measure of model fit that takes both the number of parameters and the number of observations into account, and measures both model fit and model complexity. Again, lower values are preferred. The BIC is calculated as Equation 12:

$$BIC = Kln(n) - 2ln(\hat{L}) \quad (13)$$

5.5. Results

5.5.1. Pilot launch

During the initial pilot launch, 118 respondents completed the survey. The mean (median) time to complete was 23 minutes and 54 seconds (18 minutes and 42 seconds) minutes, with a minimum of 3 minutes and 24 seconds, and a maximum time of over an hour. The initial model tested on the DCE data from the pilot sample indicated that the majority of the dimensions had evidence of monotonicity of coefficient levels – critical for the development of utility scales. The key model performance indicators were as expected and were based on the full pilot sample including people who completed the survey in 3 minutes and 24 seconds or longer. From this it was judged that data obtained from this range of completion times would produce a valid data from which to model preferences. Therefore, a minimum completion time of greater than 3 minutes was set. Three minutes was set as the minimum completion time to exclude responses from people who completed the survey very quickly. This meant that no changes were made to the study design following the pilot, and the sample was retained as part of the main study dataset.

Regarding the usability questions, it was found that only 13% of the sample agreed that they found the task difficult, 17% agreed that it was difficult to imagine the scenarios and 13% agreed that it was difficult to tell the difference between the descriptions. The free-text question did not indicate any concerning issues. This evidence was used to support the choice set formats used and launch the full sample data collection.

5.5.2. Sample – Completion process and time taken

Overall, 1,226 online panel members accessed the survey. Of these, 1,177 (96.0%) consented, 175 (14.3%) dropped out during the survey, 76 (6.2%) completed the survey in less than three minutes, and 975 (79.5%) fully completed the full survey in more than three minutes (this included the 118 respondents from the pilot launch). The mean (median) completion time was 26 minutes and 24 seconds (22 minutes and 12 seconds). **Figure 26** reports the time taken per task in seconds. A large majority of the tasks (68.6%) were completed in 30 seconds of less, and the mean (median) completion time was 35 (17). Overall, the 40 blocks of tasks were completed between 17 and 32 times, and the 20 blocks (obtained by combining the measure order blocks) were completed between 34 and 57 times (so there are between 34 and 57 observations for each choice set in the design). The results of the usability questions are displayed in **Figure 27**. The majority of respondents agreed, or strongly agreed that the tasks were not difficult.

Figure 26: Time taken per task

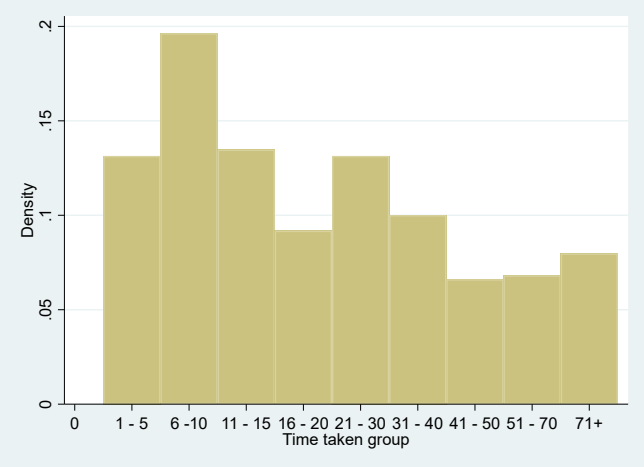
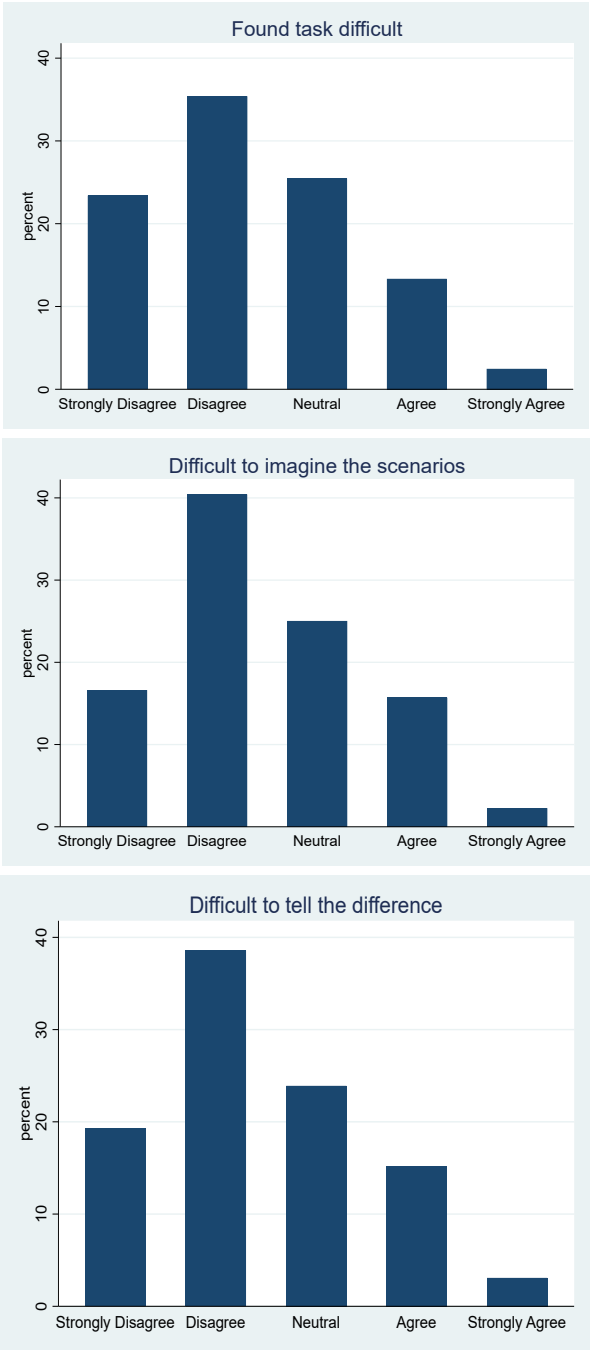


Figure 27: Frequency charts of respondent reported usability questions



5.5.3. Sample – Demographics

Table 41 reports the demographic characteristics of the sample in comparison to the available statistics for the Australian population for age and gender [242]. The sample was generally representative of the Australian population in age, gender and income, but respondents were more highly educated and more likely to be born in Australia. Overall, 44% self-reported a long-term health condition. The ASCOT utility scores were higher than those from the EQ-5D-5L.

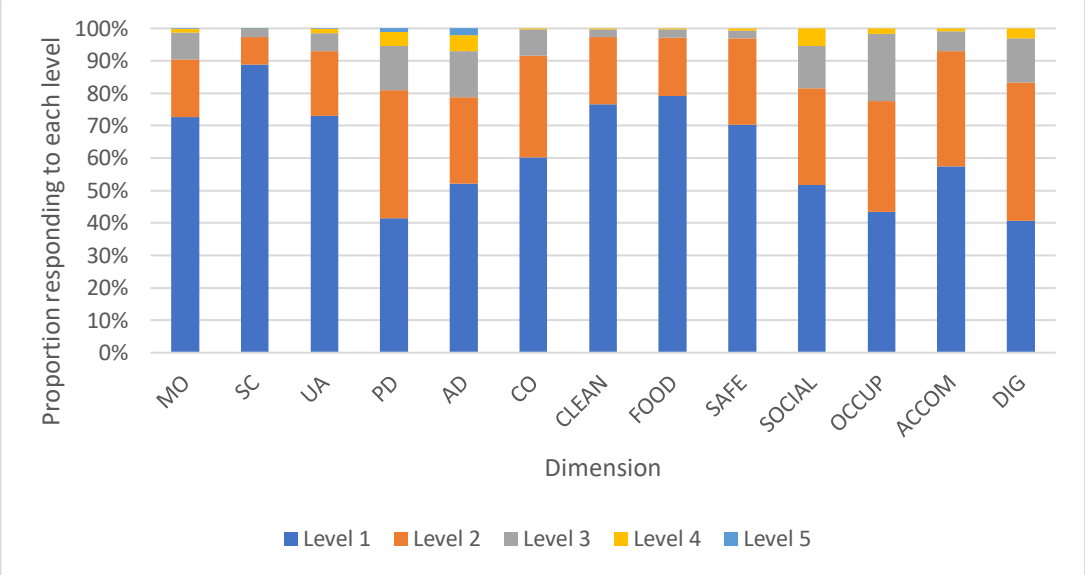
Table 41: Sample demographics

Characteristic	N(%)	Approx. Aust population
Female	495 (50.8)	51
Age Group (years)		
Mean (sd)	46.0 (16.6)	
18 – 24	115 (11.7)	12
25 – 34	176 (18.1)	18
35 – 44	182 (18.6)	19
45 – 54	176 (18.1)	18
55 – 64	148 (15.2)	15
65+	178 (18.3)	18
Marital status		
Married/partner	585 (60.0)	48
Single/widowed/separated	390 (40.0)	52
Highest education level		
Secondary school highest level	275 (28.2)	61
Further education	700 (71.2)	39
Income (Australian $)[a]		
0 to 80,000	643 (75.8)	70
80,001 +	205 (24.2)	30
Prefer not to say	127 (13.0)	
Country of birth		
Australia	748 (76.7)	67
Other	227 (23.3)	33
Number of children		
0	554 (56.8)	N/A
1	141 (14.5)	N/A
2	163 (16.7)	N/A
3+	117 (12.0)	N/A
Health status		
Excellent	107 (11.0)	N/A
Very good	343 (35.2)	N/A
Good	339 (34.8)	N/A
Fair	154 (15.8)	N/A
Poor	32 (3.3)	N/A
Has long-term condition	431 (44.2)	N/A
EQ-5D-5L utility score (m(sd))	0.773 (0.23)	N/A
ASCOT utility score (m(sd))	0.846 (0.16)	N/A
Hospitalised in last 12 months	257 (26.4)	N/A
Ever experienced serious illness:		
In self	285 (29.2)	N/A
In family	427 (43.8)	N/A
In caring for others	213 (21.9)	N/A

[a] Australian dollar = 0.67 US dollars as of July 2018; sd: Standard Deviation; EQ-5D-5L utility calculated using Norman et al (2013); ASCOT utility calculated using Netten et al (2013); Demographics taken from the Australian Bureau of Statistics; N/A: Not available

Figure 28 displays the proportion of respondents answering at each level of each dimension of each instrument. The respondents are distributed primarily across the first four severity levels of the EQ-5D-5L and the first three of the ASCOT dimensions. Few respondents answered at the most severe level of each dimension. This is expected given that the sample was recruited from the general population.

Figure 28: Percentage of respondents answering at each level of each dimension



5.5.4. Conditional Logit models

Table 42 reports the results for two models, one where the coefficients are unrestricted, and a second where any disordered coefficients are constrained to be ordered within dimensions. In **Model 15**, all responses across both versions of the survey are pooled (that is regardless of the order of the instrument). Non-monotonic coefficients are highlighted in bold, categorisation of p-values for the difference between the coefficient estimates and the omitted baseline level (1) are indicated by stars, and actual p-values for the significance between adjacent levels (relative to the immediately better level) are reported in the 'sig (btwn)' column.

The rank of the dimension (using the overall magnitude of the disutility at the worst level as an indicator of the dimensions on which the most weight is placed) is also reported. The coefficient with the largest decrement is pain/discomfort and mobility and the coefficient with the smallest estimate is social participation. For **Model 15**, the magnitude of overall dimension level coefficients can be used to understand the relative weight placed on dimensions. The dimensions with the largest disutility include four dimensions from the EQ-5D-5L (pain/discomfort, mobility, anxiety/depression, self-care) and one from the ASCOT (control).

This is followed by two from the ASCOT (food and drink, safety), EQ-5D-5L usual activities and another two ASCOT dimensions (cleanliness, occupation). The remaining three ASCOT dimensions (accommodation, dignity, social participation) have smaller coefficients. The magnitude of the estimates suggests that dimensions from one of the instruments were not consistently preferred to dimensions from the other. There is some evidence of non-significant disordered levels for three dimensions (usual activities levels 4 (severe) and 5 (unable to), pain/discomfort levels 2 (slight) and 3 (moderate), and accommodation levels 1 (home is as clean and comfortable as I want) and 2 (home is adequately clean and comfortable)). The EQ-5D-5L dimension coefficients increase significantly between levels 3 (moderate) and 4 (severe), and for three of the dimensions (mobility, pain/discomfort, anxiety/depression) the difference between levels 4 (severe) and 5 (extreme/unable to) was also significant. For the ASCOT, the difference between levels 3 and 4 (the two most severe levels with different severity descriptors used for each) for all eight dimensions was significant.

Model 16 is an adaptation of **Model 15**, where the disordered coefficients are constrained to impose ordering. This means that increases in severity result in a decrease or no change in utility (rather than an increase), and mimics a model that could be used for estimating a value set. Ordering was imposed on four pairs of levels, UA4 and UA5 (which now both result in a decrement of -0.276), PD2 and PD3 (-0.264), OC2 and OC3 (-0.077) and AC1 and AC2 (0).

Table 42: Conditional logit models for the overall sample

Parameter	Model 15: Overall model				Model 16: Model 1 ordered			
	Coef. (p)[a]	SE[b]	Sig (btwn)[c]	Rank[d]	Coef. (p)	SE	Sig (btwn)	Rank
MO2	-0.112*	0.046	0.015	2	-0.111*	0.046	0.016	2
MO3	-0.246***	0.046	0.005		-0.245***	0.046	0.004	
MO4	-0.599***	0.045	<0.001		-0.601***	0.045	<0.001	
MO5	-0.799***	0.049	<0.001		-0.797***	0.049	<0.001	
SC2	0.000	0.044	0.870	5	-0.001	0.043	0.984	5
SC3	-0.196***	0.045	<0.001		-0.196***	0.045	<0.001	
SC4	-0.479***	0.045	<0.001		-0.479***	0.044	<0.001	
SC5	-0.516***	0.046	0.433		-0.517***	0.045	0.417	
UA2	-0.019	0.048	0.620	9	-0.022	0.048	0.653	9
UA3	-0.024	0.047	0.925		-0.025	0.047	0.941	
UA4	-0.278***	0.048	<0.001		-0.276***	0.042	<0.001	
UA5	**-0.271*** **	0.047	0.875		-0.276***	0.042	n/a	
PD2	-0.275***	0.048	<0.001	1	-0.264***	0.042	<0.001	1
PD3	**-0.256*** **	0.050	0.705		-0.264***	0.042	n/a	
PD4	-0.694***	0.048	<0.001		-0.693***	0.048	<0.001	
PD5	-0.848***	0.046	0.002		-0.848***	0.046	0.002	
AD2	-0.038	0.046	0.342	3	-0.042	0.045	0.352	3
AD3	-0.199***	0.048	0.001		-0.203***	0.048	<0.001	
AD4	-0.574***	0.047	<0.001		-0.578***	0.047	<0.001	
AD5	-0.710***	0.048	0.004		-0.717***	0.047	0.003	
CO2	-0.160***	0.042	<0.001	4	-0.158***	0.042	<0.001	4
CO3	-0.247***	0.042	0.035		-0.246***	0.042	0.032	
CO4	-0.667***	0.041	<0.001		-0.668***	0.041	<0.001	
CL2	-0.111**	0.044	0.015	8	-0.112**	0.044	0.011	8
CL3	-0.188***	0.043	0.077		-0.188***	0.043	0.080	
CL4	-0.294***	0.044	0.010		-0.295***	0.044	0.009	
FD2	-0.102*	0.045	0.035	6	-0.106*	0.044	0.017	6
FD3	-0.259***	0.043	<0.001		-0.261***	0.043	<0.001	
FD4	-0.361***	0.046	0.033		-0.362***	0.046	0.033	
SA2	-0.058	0.043	0.160	7	-0.059	0.043	0.173	7
SA3	-0.127**	0.042	0.111		-0.126**	0.042	0.121	
SA4	-0.330***	0.046	<0.001		-0.328***	0.042	<0.001	
SP2	-0.009	0.043	0.816	13	-0.008	0.043	0.843	13
SP3	-0.046	0.045	0.410		-0.047	0.045	0.379	
SP4	-0.146***	0.041	0.017		-0.147***	0.041	0.017	
OC2	-0.080	0.046	0.108	10	-0.077*	0.039	0.048	10
OC3	**-0.072**	0.044	0.861		-0.077*	0.039	n/a	
OC4	-0.234***	0.043	<0.001		-0.235***	0.043	<0.001	
AC2	**0.035**	0.043	0.437	11	0	n/a[e]	n/a	11
AC3	-0.055	0.041	0.028		-0.073	0.035	0.036	
AC4	-0.199***	0.044	<0.001		-0.218***	0.037	<0.001	
DI2	-0.006	0.041	0.880	12	-0.004	0.040	0.917	12
DI3	-0.064	0.044	0.162		-0.064	0.044	0.150	
DI4	-0.159***	0.042	0.023		-0.162***	0.042	0.018	
No obs[f]	14,625				14,625			
LL[g]	-8,949				-8,949			
AIC[h]	17,986				17,979			
BIC[i]	18,351				18,310			

[a] Coefficient estimate; [b] standard error; [c] significance between adjacent levels relative to the immediately better level; [d] rank defined by the magnitude of the worst level; [e] not applicable; [f] Number of observations; [g] Log-likelihood; [h] Akaike Information Criterion; [g] Bayesian Information Criterion p-values for the difference between the coefficient and baseline indicated by stars: *** 0.001, ** 0.01; *0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

5.5.5. Models including interactions

Table 43 reports the models including interaction terms. **Model 17** includes the interactions that are included when SCRQoL problems are high, but HRQoL is no problem (to test whether having a range of SCRQoL concerns is perceived as more manageable when HRQoL is not problematic). To allow for a comparison of the magnitude of these coefficients, an interaction that is active when a health state includes any ASCOT dimension at the worst level is also included (ASCOT N4 term). Only the estimate for MO1 x ASCOT N4 is a (non-significant) moderating coefficient meaning that utility is improved when a scenario has problems with SCRQoL, but combined with no problems in mobility (so indicating that the SCRQoL issues perceived as slightly less bad). The largest further decrement is the ASCOT N4 term, which indicates that having at least one severe issue with SCRQoL results in a further decrease in utility. As a result of the inclusion of the interaction terms, the standard error of the EQ-5D-5L main effects estimates is substantially larger than the ASCOT main effect estimates.

Model 18 displays the interactions included when HRQoL problems (including both severe and extreme levels) are high, but there are no SCRQoL issues (to test whether having HRQoL problems is perceived as less problematic when no SCRQoL issues are experienced). For comparison, a term that is included when any EQ-5D-5L dimension is at level 4 or 5, irrespective of the ASCOT level, is also estimated (the EQ-5D-5L N45 term). There are one significant and three non-significant moderating interactions. The significant moderating interaction suggests that having severe or extreme HRQoL problems is less of an impact if personal cleanliness and comfort is not an issue. The non-significant moderators relate to severe EQ-5D-5L problems, and no SCRQoL issues in access to food and drink, social participation, and accommodation. The EQ-5D-5L N45 term results in an extra decrement, but is not as large as three of the other interactions resulting in a decrement. Appendix 13 (**Model 51**) reports the full set of interactions combined in a single model. The pattern of coefficient estimates is similar, but the AIC and BIC results for this model are lower.

Table 43: Exploratory analysis of interactions

Parameter	M 17: Interactions of HRQoL/poor SCRQoL		M 18: Interactions of SCRQoL/poor HRQoL	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE
Main Effects				
MO2	-0.013	0.142	-0.121**	0.046
MO3	-0.135	0.139	-0.254***	0.047
MO4	-0.493***	0.139	-0.601***	0.046
MO5	-0.702***	0.140	-0.815***	0.051
SC2	-0.122*	0.122	-0.008	0.044
SC3	-0.331**	0.124	-0.224***	0.046
SC4	-0.609***	0.127	-0.493***	0.045
SC5	-0.641***	0.126	-0.524***	0.047
UA2	-0.016	0.139	-0.034	0.049
UA3	-0.030	0.140	**-0.004**	0.048
UA4	-0.285*	0.142	-0.262***	0.050
UA5	**-0.277****	0.142	-0.271***	0.048
PD2	-0.283	0.117	-0.284***	0.050
PD3	**-0.269**	0.118	**-0.280*****	0.051
PD4	-0.707	0.117	-0.692***	0.050
PD5	-0.857	0.118	-0.848***	0.047
AD2	-0.064	0.141	-0.046	0.047
AD3	-0.237*	0.141	-0.193***	0.050
AD4	-0.600***	0.137	-0.570***	0.050
AD5	-0.727***	0.141	-0.717***	0.050
CO2	-0.168***	0.042	-0.537***	0.135
CO3	-0.256***	0.043	-0.660***	0.136
CO4	-0.661***	0.042	-1.041***	0.136
CL2	-0.120**	0.045	**0.184**	0.129
CL3	-0.256***	0.044	**0.110**	0.130
CL4	-0.277***	0.045	**0.004**	0.130
FD2	-0.106*	0.046	-0.067	0.141
FD3	-0.266***	0.044	-0.206	0.139
FD4	-0.343***	0.047	-0.304	0.139
SA2	-0.054	0.044	-0.076	0.150
SA3	-0.130**	0.043	-0.143	0.151
SA4	-0.312***	0.043	-0.352**	0.147
SP2	-0.013	0.043	0.249	0.150
SP3	-0.033	0.046	0.213	0.151
SP4	-0.124**	0.043	0.103	0.147
OC2	-0.080*	0.046	-0.177	0.125
OC3	-0.085	0.045	-0.179	0.123
OC4	-0.224***	0.043	-0.338**	0.123
AC2	**0.035**	0.044	0.080	0.173
AC3	-0.046	0.042	-0.024	0.172
AC4	-0.184***	0.045	-0.145	0.171
DI2	-0.011	0.042	-0.131	0.131
DI3	-0.058	0.045	-0.188	0.132
DI4	-0.145***	0.043	-0.264*	0.128
Interactions				
ASCOT Level 4	-0.146	0.096		
MO1 x ASCOT Level 4	0.101	0.141	n/a	n/a
SC1 x ASCOT Level 4	-0.133*	0.128	n/a	n/a
UA1 x ASCOT Level 4	-0.010	0.143	n/a	n/a
PD1 x ASCOT Level 4	-0.025	0.120	n/a	n/a
AD1 x ASCOT Level 4	-0.026	0.142	n/a	n/a
EQ-5D Levels 4/5	n/a[c]	n/a	-0.057	0.100
CO1 x EQ-5D Levels 4/5	n/a	n/a	-0.410**	0.137
CL1 x EQ-5D Levels 4/5	n/a	n/a	0.321**	0.129
FD1 x EQ-5D Levels 4/5	n/a	n/a	0.043	0.142
SA1 x EQ-5D Levels 4/5	n/a	n/a	-0.005	0.128
SP1 x EQ-5D Levels 4/5	n/a	n/a	0.275	0.152
OC1 x EQ-5D Levels 4/5	n/a	n/a	-0.101	0.134
AC1 x EQ-5D Levels 4/5	n/a	n/a	0.061	0.171
DI1 x EQ-5D Levels 4/5	n/a	n/a	-0.134	0.132
No Obs[d]	14,625		14,625	
LL[e]	-8,944		-8,938	
AIC[f]	17,988		17,982	
BIC[g]	18,402		18,421	

5.5.6. Scale testing across subsamples

Regarding measure order (**Table 44**), **Model 21** reports the coefficients from the restricted model, and the scale parameter controlling for the presentation order. Comparing the pooled model with the two unrestricted model for each order separately gives an LR statistic of 118 (greater than the critical value from a Chi Square distribution, with 46 degrees of freedom, of 61.7), and the scale parameter is significant at the 0.001 level, hence the null hypothesis of preference homogeneity is rejected. This indicates that the order of measure presentation has an impact on the results. The EQ-5D-5L dimensions have a larger decrement when presented first (**Model 19**). When ASCOT is presented first (**Model 20**), the magnitude of the disutility of control becomes larger in comparison to the EQ-5D-5L dimensions, and there is increased non-monotonicity, in particular for the EQ-5D-5L dimensions. Non-monotonicity occurred in only one dimension of the EQ-5D-5L (anxiety/depression) in **Model 19**, but occurred in three (across self-care, usual activities, pain/discomfort) in **Model 20**. The overall pattern of ranking of the coefficients is reasonably similar. Dimensions ranked one to five in **Model 19** are also ranked one to five in **Model 20** (with four of the coefficients in a different order). Dimensions ranked six to eight are also ranked six to eight (in a different order) and those ranked nine to 12 are in the same order across both models. Therefore ten of the 13 estimates are ranked in the same order across the models. In contrast to the overall models reported in Section 5.5.4, generating ordered models here is not important as the analysis is conducted to assess disordering and the impact of measure presentation order.

Appendix 14 reports the full models for the scale testing across demographic variables. For gender the LR statistic was 49.97 (**Table 76**) reports the models, for age was 93.69 (**Table 77**), and for condition status was 53.25 (**Table 78**). Therefore, the null hypothesis of scale homogeneity is accepted for gender and condition status, but not for age.

Table 44: Conditional logit and heteroskedastic pooled models by measure order

Parameter	Model 19: EQ-5D-5L appearing first				Model 20: ASCOT appearing first				Model 21: Restricted pooled	
	Coef. (p)[a]	SE[b]	Sig (btwn)[c]	Rank[d]	Coef. (p)	SE	Sig (btwn)	Rank	Coef. (p)	SE
MO2	-0.195**	0.067	0.003	1	-0.041	0.065	0.521	4	-0.102*	0.044
MO3	-0.417***	0.068	0.001		-0.098	0.064	0.390		-0.227***	0.045
MO4	-0.795***	0.065	<0.001		-0.424***	0.062	<0.001		-0.561***	0.047
MO5	-1.061***	0.072	<0.001		-0.566***	0.068	0.025		-0.747***	0.053
SC2	-0.044	0.063	0.404	4	**0.044**	0.061	0.534	5	0.001	0.041
SC3	-0.193***	0.065	0.034		-0.189**	0.062	<0.001		-0.187***	0.043
SC4	-0.465***	0.063	<0.001		-0.501***	0.062	<0.001		-0.457***	0.043
SC5	-0.545***	0.066	0.233		**-0.496*** **	0.064	0.939		-0.489***	0.045
UA2	-0.039	0.070	0.560	9	0.000	0.067	0.917	9	-0.017	0.046
UA3	-0.069	0.068	0.636		**0.017**	0.066	0.788		-0.020	0.045
UA4	-0.206**	0.070	0.034		**-0.349*** **	0.067	<0.001		-0.267***	0.046
UA5	-0.261***	0.068	0.397		**-0.281*** **	0.065	0.272		-0.258***	0.045
PD2	-0.195**	0.059	0.006	2	-0.348***	0.068	<0.001	1	-0.265***	0.046
PD3	-0.228**	0.072	0.644		**-0.281*** **	0.070	0.338		-0.245***	0.048
PD4	-0.612***	0.070	<0.001		-0.779***	0.068	<0.001		-0.663***	0.049
PD5	-0.787***	0.066	0.015		-0.908***	0.064	0.065		-0.807***	0.048
AD2	**0.014**	0.066	0.996	3	-0.087	0.064	0.191	3	-0.039	0.043
AD3	-0.164*	0.070	0.008		-0.234***	0.068	0.025		-0.189***	0.046
AD4	-0.440***	0.069	<0.001		-0.703***	0.066	<0.001		-0.550***	0.047
AD5	-0.613***	0.069	0.011		-0.812***	0.067	0.094		-0.679***	0.048
CO2	-0.122*	0.061	0.038	5	-0.202***	0.058	0.001	2	-0.155***	0.040
CO3	-0.164**	0.061	0.482		-0.333***	0.059	0.023		-0.238***	0.041
CO4	-0.503***	0.060	<0.001		-0.834***	0.058	<0.001		-0.640***	0.042
CL2	-0.191**	0.064	0.004	8	-0.033	0.062	0.604	=6	-0.102*	0.042
CL3	-0.246***	0.062	0.381		-0.133*	0.060	0.102		-0.176***	0.041
CL4	-0.265***	0.064	0.749		-0.318***	0.062	0.002		-0.280***	0.043
FD2	-0.128	0.065	0.059	6	-0.080	0.062	0.260	8	-0.095*	0.043
FD3	-0.258	0.062	0.034		-0.266***	0.060	0.002		-0.247***	0.041
FD4	-0.428	0.068	0.014		-0.307***	0.065	0.542		-0.341***	0.045
SA2	-0.040	0.063	0.525	7	-0.084	0.061	0.134	=6	-0.056	0.041
SA3	-0.112	0.061	0.253		-0.153**	0.059	0.249		-0.122**	0.040
SA4	-0.357***	0.061	<0.001		-0.318***	0.059	0.005		-0.313***	0.041
SP2	-0.060	0.062	0.350	13	**0.042**	0.060	0.490	13	-0.007	0.041
SP3	-0.100	0.066	0.529		**0.007**	0.063	0.573		-0.041	0.043
SP4	**-0.085**	0.060	0.809		-0.198***	0.058	0.001		-0.140***	0.040
OC2	-0.048	0.066	0.507	10	-0.116	0.064	0.089	10	-0.078	0.044
OC3	-0.102	0.063	0.378		**-0.050**	0.061	0.268		-0.068	0.042
OC4	-0.201***	0.062	0.102		-0.272***	0.060	<0.001		-0.224***	0.041
AC2	**0.044**	0.062	0.451	11	**0.029**	0.060	0.688	11	0.033	0.041
AC3	**0.015**	0.059	0.622		-0.119*	0.058	0.009		-0.055	0.039
AC4	-0.157**	0.063	0.003		-0.240***	0.061	0.032		-0.191***	0.042
DI2	-0.018	0.059	0.792	12	**0.005**	0.056	0.972	12	-0.004	0.038
DI3	**0.002**	0.064	0.741		-0.128**	0.062	0.024		-0.063	0.042
DI4	-0.114	0.061	0.051		-0.206***	0.059	0.176		-0.151***	0.040
Order									0.094*	0.047
No obs[e]	6,975				7,650				14,625	
LL[f]	-4,280				-4,608				-8,947	
AIC[g]	8,649				9,304				17,984	
BIC[h]	8,981				9,640				18,357	

[a] Coefficient estimate; [b] standard error; [c] significance between adjacent levels relative to the immediately better level; [d] rank defined by the magnitude of the worst level; [e] Number of observations; [f] Log-likelihood; [g] Akaike Information Criterion; [h] Bayesian Information Criterion p-values for the difference between the coefficient and baseline indicated by stars: *** 0.001, ** 0.01; *0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

5.5.7. Sensitivity analysis - Time taken by task and overall

The results of the time taken sensitivity analyses are reported in Appendix 15. The results of the assessment of model ordering based on both task and overall respondent level exclusions suggests that completing in a shorter and longer time overall contributes to increased disordering between coefficient levels within certain dimensions, although the pattern is not consistent.

The results of the scale testing differ depending on whether the sample is split by time taken per task or overall. For the time taken per task, the null hypothesis of preference homogeneity was rejected, and therefore scale differs according to the time taken to complete the tasks. For the time taken overall, that the null hypothesis of preference homogeneity was accepted, meaning that the scale does not differ between samples.

5.5.8. Assessing heterogeneity – Latent class

For the latent class analysis, the two class model (**Model 22**) displayed the lowest BIC (18,093) of the models including up to six classes (see **Table 45**). The resulting estimates are shown in **Table 46**. The first class includes 55% of the sample who demonstrate strong ordered preferences across ten of the EQ-5D-5L and ASCOT dimensions. The second class (45%) includes a less clear pattern of preferences, with evidence of disordering across dimension levels. Those in class one are more likely to be aged over 60 and have a long-term condition than those in class two. For comparison, the three to six class models are included in Appendix 16. Interpretation of the class structure is more complex which would be expected, as these models do not explain the data to the same level as the two and three class models.

Table 45: Latent class model performance statistics

Model and Class number	LL	BIC	AIC
2 class **(Model 22)**	**-8,730**	**18,093**	17,644
3 class **(Model 67)**	-8,607	18,178	17,495
4 class (**Model 68**)	-8,496	18,286	17,368
5 class (**Model 69**)	-8,429	18,482	17,330
6 class **(Model 70)**	-8,343	18,641	**17,254**

Bold values: Best model indicator

Table 46: Two class latent class model

	Model 22: Two class model	
Parameter	**Class 1**	**Class 2**
MO2	-0.085	-0.100
MO3	-0.383	-0.135
MO4	-0.807	-0.440
MO5	-1.098	-0.592
SC2	**0.048**	-0.025
SC3	-0.269	-0.162
SC4	-0.879	-0.251
SC5	-1.151	**-0.054**
UA2	-0.044	**0.010**
UA3	-0.052	-0.005
UA4	-0.505	-0.099
UA5	-0.572	**-0.028**
PD2	-0.433	-0.191
PD3	**-0.425**	**-0.176**
PD4	-1.382	-0.197
PD5	-1.541	-0.373
AD2	-0.134	**0.001**
AD3	-0.420	-0.065
AD4	-1.439	**0.092**
AD5	-1.590	-0.086
CO2	-0.228	-0.126
CO3	-0.234	-0.254
CO4	-1.102	-0.298
CL2	-0.264	-0.043
CL3	-0.408	-0.079
CL4	-0.547	-0.154
FD2	-0.068	-0.150
FD3	-0.378	-0.181
FD4	-0.435	-0.327
SA2	-0.234	**0.063**
SA3	-0.410	**0.070**
SA4	-0.693	-0.077
SP2	-0.118	**0.073**
SP3	-0.118	**0.012**
SP4	-0.369	**0.042**
OC2	-0.157	-0.057
OC3	-0.191	**-0.018**
OC4	-0.473	-0.099
AC2	**0.025**	**0.038**
AC3	-0.141	**0.016**
AC4	-0.315	-0.113
DI2	-0.039	**0.000**
DI3	-0.183	**0.018**
DI4	-0.361	-0.004
Demographic		
Age Cat (18-60 and 60+)	0.758	0.000
Gender	0.746	0.000
Has Long-term Condition	0.192	0.000
Class Share	0.550	0.450
N Obs[a]	14,625	
LL[b]	-8,730	

[a] Number of observations; [b] Log-Likelihood; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity;

5.5.9. Assessing heterogeneity – Mixed logit

Appendix 17 reports mixed logit models including one parameter for each dimension. These models were estimated to test a range of mixed logit specifications to draw inferences about which specifications should be included in the models estimating all levels of all dimensions. The specifications tested included coefficient distribution, coefficient correlations, and number of draws. **Model 71** to **Model 72** report mixed logit models including one parameter for each dimension, with all specified as random and with normally distributed coefficients. Across both models, the standard deviations suggest that the ASCOT dimensions cleanliness and accommodation do not display evidence of heterogeneity, but the other dimensions are significantly heterogeneous. The model with increased draws is more efficient. **Model 73** reports the overall dimension level model, but with log-normally distributed parameters. The pattern of significant heterogeneity is the same. **Model 74** reports the overall dimension level model specifying that the parameters are correlated. The level of heterogeneity is increased, but the AIC and BIC are higher indicating lower model performance.

These results led to a series of decisions about the main models to report in this chapter. First, given the difficulty in fitting log-normal models with parameters that are inconsistent in conditional logit (as is the case with the baseline model reported here), the dimension level models reported below are specified with normally distributed parameters. Second, the difficulty in specifying which dimension levels are correlated across the 44 parameters, and the complexity of the modelling approach, it was decided to focus on models reporting uncorrelated coefficients. Third, the models reported in the main text increase the number of draws to 1,000, and the burn rate to 44, with the parameters specified as random.

Table 47 reports the most complex models with all 44 EQ-5D-5L and ASCOT dimension level parameters specified as random (**Model 23**), and normally distributed uncorrelated coefficients. The model (**Model 75**) including a lower number of draws and burn rates is included in Appendix 17. The AIC and BIC are lower for **Model 23** which suggests that increasing the number of draws, burn rate, and specifying a difficult maximisation procedure increases the performance of the model. The results suggest that the most severe level of the majority of the dimensions has at least some evidence of significant preference heterogeneity across both model specifications. There are some differences in the magnitude of the standard deviations between the models, but the direction of the differences is not consistent.

Table 47 also reports **Model 24** where the EQ-5D-5L dimension levels are specified as random (Appendix 17 reports the version with less draws (**Model 76**)). The MO, SC, PD and AD dimensions have a higher level of heterogeneity than UA, particularly at the more severe dimension levels. The AIC and BIC again suggest that the model with increased draws and a higher burn rate performs better.

Table 47 reports **Model 25** where the ASCOT dimensions are random, and the EQ-5D-5L dimensions are fixed (Appendix 17 reports the version with lower draws (**Model 77**)). Across both models, there is evidence of heterogeneity at the more severe levels of control, social, occupation, accommodation and dignity. The models including more draws are more conservative as they demonstrate less evidence of heterogeneity, and it also has a lower AIC and BIC. Given the more conservative nature of the models increasing the number of draws, and specifying a difficult maximisation procedure, and the complex nature of the models this specification was used for the further exploratory models tested with different sets of random parameters based on past valuation work, and the results of the overall model including all parameters as random.

Appendix 17 reports exploratory mixed logit models based on preference patterns observed in the literature and extracting sets of parameters from the most complex model with all EQ-5D and ASCOT dimension levels specified as random. **Model 78** specifies the most severe dimension levels as random in line with other valuation work suggesting increased variation in preferences around more severe health states [172]. There is significant heterogeneity for the majority of the most severe dimension levels, with the ASCOT dimensions accommodation, safety and cleanliness exhibiting homogeneity. **Model 79** includes as random those terms that had the most significant standard deviations from **Model 23** which includes all 44 parameters as random. There are differences in comparison to that model, as a number of the standard deviations are no longer significant. This suggests that the pattern of heterogeneity is dependent on the sets of parameters specified as random. **Model 80** demonstrates that the ten most heterogeneous dimensions are consistently significant across both models. This model has the lowest AIC and BIC of the exploratory models. **Model 81** demonstrates that the five most heterogeneous parameters retain strong evidence of heterogeneity when tested on their own. This suggests that the ten most heterogeneous parameters have strong evidence of heterogeneity. Those parameters with less significant heterogeneity differ in terms of the level of heterogeneity across different models.

Table 47: Mixed Logit models – EQ-5D-5L and ASCOT combinations as random

	M 23: All dimension		M 24: EQ-5D-5L vary		M 25: ASCOT vary	
Parameter	**Coef.** [a]	**SD**	**Coef.**	**SD**	**Coef.**	**SD**
MO2	-0.170**	0.131	-0.120***	0.073	-0.141**	N/A[b]
MO3	-0.348***	0.096	-0.259***	0.007	-0.279***	N/A
MO4	-0.834***	0.347**	-0.666***	0.365***	-0.678***	N/A
MO5	-1.112***	0.841***	-0.911***	0.709***	-0.885***	N/A
SC2	-0.024	0.000	-0.010	0.041	-0.011	N/A
SC3	-0.279***	0.431**	-0.215***	0.351**	-0.226***	N/A
SC4	-0.659***	0.540***	-0.542***	0.454***	-0.531***	N/A
SC5	-0.725***	0.795***	-0.584***	0.575***	-0.565***	N/A
UA2	-0.059	0.309*	-0.036	0.052	-0.039	N/A
UA3	**-0.036**	0.248	**-0.030**	0.098	**-0.036**	N/A
UA4	-0.398***	0.276	-0.328***	0.249*	-0.314***	N/A
UA5	**-0.387*** **	0.281*	**-0.315*** **	0.089	**-0.307*** **	N/A
PD2	-0.355***	0.437***	-0.308***	0.171	-0.299***	N/A
PD3	-0.340***	0.120	**-0.288*** **	0.085	**-0.274*** **	N/A
PD4	-0.939***	0.585***	-0.784***	0.462***	-0.750***	N/A
PD5	-1.173***	0.750***	-0.970***	0.576***	-0.921***	N/A
AD2	-0.055	0.374**	-0.029***	0.316***	-0.054	N/A
AD3	-0.290***	0.404**	-0.229***	0.145	-0.230***	N/A
AD4	-0.816***	0.775***	-0.640***	0.499***	-0.651***	N/A
AD5	-1.011***	0.854***	-0.806***	0.429***	-0.787***	N/A
CO2	-0.183***	0.171	-0.174***	N/A	-0.144**	0.063
CO3	-0.333***	0.314*	-0.284***	N/A	-0.255***	0.265*
CO4	-0.903***	0.842***	-0.735***	N/A	-0.744***	0.698***
CL2	-0.156**	0.067	-0.128**	N/A	-0.119*	0.0202
CL3	-0.258***	0.401***	-0.221***	N/A	-0.284***	0.264*
CL4	-0.393***	0.221	-0.332***	N/A	-0.400***	0.028
FD2	-0.138*	0.050	-0.111*	N/A	-0.065	0.007
FD3	-0.342***	0.318*	-0.288***	N/A	-0.284***	0.136
FD4	-0.490***	0.437***	-0.400***	N/A	-0.400***	0.268
SA2	-0.071	0.312*	-0.069	N/A	-0.065	0.211
SA3	-0.172**	0.009	-0.137**	N/A	-0.139**	0.025
SA4	-0.443***	0.322*	-0.365***	N/A	-0.361***	0.218
SP2	-0.012	0.383***	0.016	N/A	-0.011	0.193
SP3	-0.074	0.047	0.052	N/A	-0.056	0.113
SP4	-0.212***	0.387***	0.181***	N/A	-0.158***	0.261*
OC2	-0.094	0.214	0.082	N/A	-0.088	0.087
OC3	-0.110	0.431***	0.079	N/A	-0.089	0.308**
OC4	-0.322***	0.567***	0.253***	N/A	-0.268***	0.472***
AC2	**0.032**	0.045	0.047	N/A	**0.029**	0.065
AC3	-0.081	0.264	0.059	N/A	-0.072	0.260**
AC4	-0.277***	0.260*	0.213***	N/A	-0.231***	0.208
DI2	-0.017	0.193	0004	N/A	-0.011	0.160
DI3	-0.074	0.094	0.061	N/A	-0.076	0.004
DI4	-0.214***	0.385***	0.170***	N/A	-0.177***	0.276**
N	14,625		14,625		14,625	
LL[c]	-8,811		8,887		-8,906	
AIC[d]	17,799		17,865		17,948	
BIC[e]	18,528		18,263		18,511	

[a] Coefficient estimate; [b] not applicable; [c] Log-Likelihood; [d] Akaike Information Criterion; Bayesian Information Criterion; p-values for difference between coefficient and baseline ***0.001, **0.01; *0.05; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

5.5.10. Assessing heterogeneity – Generalised Multinomial Logit Model

Table 48 reports a range of GMNL models incorporating both preference and scale heterogeneity and specifying that the parameters are uncorrelated. Given the results from the mixed logit regarding the estimation of more conservative models with better performance indicators using increased draws, burn rate, and specifying difficult maximisation, this specification was used for all GMNL models. Given the difficulties in estimating dimension level models with log-normal distributions, random parameters with normal distributions were specified throughout the GMNL analysis. Again, the focus was also on modelling uncorrelated parameters.

Model 26 displays the indicators of heterogeneity across the 20 EQ-5D-5L dimension level parameters and suggests evidence of heterogeneity at the more severe levels of the MO, SC, PD and AD dimensions, which is consistent with the findings from the mixed logit analysis. **Model 27** includes the 13 ASCOT parameters with evidence of heterogeneity in the mixed logit analysis. The GMNL model suggests less evidence of heterogeneity in comparison to the mixed logit, but scale heterogeneity is apparent. The results of **Model 28** suggests preference heterogeneity across most of the severe levels of both the EQ-5D and ASCOT. **Model 29** differs to the mixed logit analysis, as eight of the 20 parameters displaying evidence of heterogeneity in the mixed logit model including all parameters as random are not heterogeneous in the GMNL model. Across all models, the Tau statistics were highly significant which indicates the presence of scale heterogeneity that there was scale heterogeneity present in the dataset. This finding supports the earlier analysis of scale differences which were significant across a number of demographic indicators.

5.6. Discussion

5.6.1. Summary and explanation of findings

This study has investigated the relative magnitude of health and social care related QoL dimensions included in the EQ-5D-5L and ASCOT using a stated preference approach. The results found that respondents make choices reflecting their trade-offs between diverse dimensions of QoL within an overall utility framework combining preferences for certain aspects of HRQoL and SCRQoL. The estimates demonstrate that the HRQoL and SCRQoL outcomes included in the EQ-5D-5L and ASCOT have different levels of importance (using coefficient size as a proxy for importance) to a large and generally representative pool of respondents. The magnitude of

preferences is generally higher for EQ-5D-5L HRQoL dimensions, with some exceptions, and there is evidence of preference heterogeneity.

This work adds preference evidence to the evidence of on the measurement relationship between the measures reported in Chapter 4 and demonstrates that trading across diverse areas of QoL within the same valuation framework is feasible. The results have implications for the future development and valuation of measures including wider areas of QoL. They demonstrate that preferences for the wider domains of QoL cannot be separated from preferences for narrower domains valued separately (for example HRQoL). Therefore to facilitate accurate decision making, the broader domains need to be valued a part of a unified measurement framework. The results also demonstrate that DCE is a feasible approach to valuing diverse domains of QoL. The task used is relatively challenging, but the majority of the sample self-reported not finding the task difficult to complete. This may be due to a higher level of education amongst the sample (in particular those educated to degree level of higher), or due to inconsistencies in self-reporting around perceived task difficulty. To improve the inclusiveness of DCE tasks, further research could attempt to understand task completion in those with lower education levels.

Table 48: Generalised multinomial logit models for the EQ-5D-5L and ASCOT data

	Model 26: EQ-5D-5L vary		Model 27: 13 ASCOT vary		Model 28: Severe levels vary		Model 29: 20 most significant in MIXL	
Parameter	**Coef.**	**SD**	**Coef.**	**SD**	**Coef.**	**SD**	**Coef.**	**SD**
MO2	-0.128*	0.100	-0.172**	N/A	-0.138*	N/A	-0.148*	N/A
MO3	-0.348***	0.129	-0.354***	N/A	-0.366***	N/A	-0.371***	N/A
MO4	-0.742***	0.131	-0.836***	N/A	-0.879***	0.136	-0.845***	0.288**
MO5	-1.144***	0.817***	-1.127***	N/A	-1.271***	0.844***	-1.195***	0.755***
SC2	-0.031	0.060	-0.043	N/A	-0.045	N/A	-0.039	N/A
SC3	-0.308***	0.131	-0.301***	N/A	-0.339	N/A	-0.318***	0.161
SC4	-0.765***	0.481***	-0.724***	N/A	-0.793***	0.509***	-0.776***	0.486***
SC5	-0.857***	0.498***	-0.820***	N/A	-0.953***	0.721***	-0.888***	0.431***
UA2	-0.035	0.266**	-0.037	N/A	-0.072	N/A	-0.059	N/A
UA3	-0.046	0.240*	-0.063	N/A	-0.067	N/A	-0.069	N/A
UA4	-0.390***	0.018	-0.392***	N/A	-0.521***	-0.276*	-0.447***	N/A
UA5	-0.463***	0.080	-0.426***	N/A	-0.503***	0.123	-0.477***	N/A
PD2	-0.421***	0.191	-0.387***	N/A	-0.424***	N/A	-0.415***	0.239
PD3	-0.410***	0.178	-0.362***	N/A	-0.441***	N/A	-0.403***	N/A
PD4	-1.098***	0.362***	-0.997***	N/A	-1.191***	0.529***	-1.074***	0.373*
PD5	-1.269***	0.577***	-1.156***	N/A	-1.390***	0.488***	-1.279***	0.051
AD2	-0.091	0.037	-0.134*	N/A	-0.141*	N/A	-0.110	0.146
AD3	-0.294***	0.196	-0.330***	N/A	-0.379***	N/A	-0.344***	0.744**
AD4	-0.945***	0.567***	-0.935***	N/A	-1.024***	0.664***	-1.000***	0.521**
AD5	-1.123***	0.467***	-1.064***	N/A	-1.244***	0.545***	-1.134***	0.062
CO2	-0.205***	N/A	-0.150	N/A	-0.186**	N/A	-0.189***	N/A
CO3	-0.266***	N/A	-0.212***	0.263***	-0.278***	N/A	-0.244***	N/A
CO4	-0.901***	N/A	-0.824***	0.718***	-0.980***	0.603***	-0.891***	0.577***
CL2	-0.199***	N/A	-0.187***	N/A	-0.200***	N/A	-0.157**	N/A
CL3	-0.325***	N/A	-0.317***	0.202	-0.352***	N/A	-0.310***	0.059
CL4	-0.471***	N/A	-0.450***	N/A	-0.472***	-0.017	-0.460***	N/A
FD2	-0.058	N/A	-0.082	N/A	-0.088	N/A	-0.067	N/A
FD3	-0.307***	N/A	-0.301***	0.119	-0.317***	N/A	-0.305***	N/A
FD4	-0.452***	N/A	-0.436***	0.421***	-0.460***	0.343***	-0.449***	0.203*
SA2	-0.092	N/A	-0.143**	0.209	-0.129*	N/A	-0.114*	N/A
SA3	-0.195***	N/A	-0.229***	N/A	-0.237***	N/A	-0.206***	N/A
SA4	-0.499***	N/A	-0.505***	0.351***	-0.604***	0.165	-0.503***	N/A
SP2	-0.018	N/A	-0.027	0.275**	-0.057	N/A	-0.027	0.248**
SP3	-0.069	N/A	-0.060	N/A	-0.092	N/A	-0.074	N/A
SP4	-0.227***	N/A	-0.214***	0.312***	-0.304***	0.176	-0.224***	0.280***
OC2	-0.154*	N/A	-0.143*	N/A	-0.179**	N/A	-0.171**	N/A
OC3	-0.178**	N/A	-0.176**	0.265	-0.178**	N/A	-0.178**	0.213*
OC4	-0.418***	N/A	-0.381***	0.192	-0.446***	0.537***	-0.428***	0.009
AC2	**0.059**	N/A	**0.025**	N/A	**0.058**	N/A	-0.018	N/A
AC3	-0.059	N/A	-0.094	N/A	-0.059	N/A	-0.095	N/A
AC4	-0.236***	N/A	-0.271***	0.163	-0.270***	0.129	-0.268***	N/A
DI2	0.005	N/A	0.004	N/A	-0.054	N/A	0.003	N/A
DI3	-0.085	N/A	-0.099***	N/A	-0.090	N/A	-0.071	N/A
DI4	-0.218***	N/A	-0.211***	-0.000	-0.266***	0.102	-0.210***	0.157
TAU	0.926***		0.913***		0.988***		0.921***	
GAMMA	0.474***		0.310*		0.226*		0.391***	
N orbs	14,625		14,625		14,625		14,625	
LL	-8,754		-8,748		-8,728		-8,744	
AIC	17,789		17,768		17,730		17,726	
BIC	18,100		18,006		17,983		17,895	

[a] Coefficient estimate; [b] not applicable; [c] Log-Likelihood; [d] Akaike Information Criterion; Bayesian Information Criterion; p-values for difference between coefficient and baseline ***0.001, **0.01; *0.05; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

The overall coefficient magnitude is an indicator of which particular QoL aspects within the classification system respondents prefer and suggest that two measures with different perspectives can be perceived as describing a broader concept within which people trade across different aspects. This pattern of preferences has implications for decision making using a conventional QALY (focusing on HRQoL), that does not include wider areas of SCRQoL. The use of an instrument generating a value set combining HRQoL and SCRQoL would be driven more by the needs of particular groups, for example people with long-term conditions or frailty. These values could be used to assess the integration of services for people with multiple and complex conditions for whom maintaining QoL within a social care setting may be a more important consideration than improving health.

There is also evidence of preference heterogeneity across all dimensions of both instruments to different degrees, with most heterogeneity apparent at the more severe dimension levels. This means that preferences for the dimensions of both instruments differ in different groups of respondents in terms of which dimension they would most want to avoid. There were no clear patterns of heterogeneity across the overall set of dimensions. This has implications for the sensitivity of decision making, as a value set applied to data from different population groups may not accurately reflect the preferences of other population groups. Value sets representing general population preferences at the overall level are preferred by many decision makers, but population specific value sets taking into account heterogeneity of preferences could be considered for sensitivity analysis. In future work, there is also the option of collecting preference data from patient groups with relevant HRQoL and SCRQoL impacts to understand how values differ between the general population, and a potentially more informed sample.

Another factor that may impact upon our results is the wording of the dimensions included. This clearly differs between the instruments, but qualitatively the wording of the EQ-5D-5L is more consistent internally than the ASCOT as it uses severity descriptors that only differ on one level across dimensions. This difference in the wording of the descriptions and the severity levels may influence valuations and potentially mask the importance of certain domains (that are important to the respondent but the way they are worded means that they do not have the same overall severity perception). This issue may be resolved by valuing health states in context, where the entire descriptive system is presented within a task, and the levels relevant to the choice set are highlighted [165].

It would also be useful to examine the relationship between the dimensions and the wording in a systematic way. Using consistent descriptions and severity levels across the constructs could lead to further insight regarding the relationship between the different dimensions. It may also lead to the development of dimension descriptors that are quite different to those in the original instruments, but further work could test whether homogenising the wording influences both self-report and trading across dimensions. However, the amount of work required to develop new instruments and associated descriptions is extensive. The developmental work conducted for the EQ-5D-5L [78] and ASCOT [126] is comprehensive, and therefore using existing instruments is a key starting point in understanding how respondents trade across dimensions.

In the models based on the survey ordering, the relative magnitude of some of the dimensions from the instrument that appears first is increased, and the scale testing is consistent with there being different patterns of preferences across the subgroups. When the ASCOT dimensions appear first, there is evidence of increased non-monotonicity, particularly for the EQ-5D-5L dimensions. However, the overall pattern of dimension ranking for each model has similarities. This provides some evidence for an order effect that may be aggregated for the overall model. Full dimension order randomisation was not imposed in the design of this study, as in other DCEs the impact of dimension ordering has not been pronounced. It may be that the difference between this work and previous studies is that here clusters of dimensions, rather than individual ones, were reordered.

5.6.2. Comparison with other EQ-5D-5L and ASCOT value sets

It is possible to compare the disutility of the coefficients for each instrument with those from valuation studies of the EQ-5D-5L and ASCOT separately. The order of magnitude of the coefficients estimated within each individual instrument shows some consistency with other published value sets produced using a range of preference elicitation methods [29, 57, 94]. Regarding EQ-5D-5L, the order of the coefficients is similar to the Australian value set derived using DCE methods [57], with mobility, pain/discomfort and anxiety/depression having the largest decrements. The order of the ASCOT dimensions is consistent with the UK value set [126] for the dimensions with the largest and smallest overall impact on ASCOT utility. One key difference for the ASCOT preferences is the weight placed on control compared to the other dimensions, which is higher in this study. This could be linked to the preferences of the Australian population, or an effect of presenting control alongside the EQ-5D-5L dimensions, where control over life might be considered differently when presented alongside specific health

aspects.

Other work has assessed the relationship between preferences elicited for both instruments in the same study. Stevens et al [243] estimated an exchange rate between EQ-5D-3L and the ASCOT using TTO valuations of each measure separately in the same respondents. The exchange rate suggested that health outcomes as measured by the EQ-5D-3L were more valued than ASCOT outcomes, but the gradient was close to one and the intercept was also small. The study did not allow for respondents to express preferences for different dimensions measured by the instruments within the same framework, as the health and social care descriptions were valued separately. In this work the two aspects were combined and show generally higher preferences for HRQoL aspects included in EQ-5D-5L with some exceptions. The studies provide complementary information from different preference elicitation approaches about how people consider HRQoL and SCRQoL as measured by the EQ-5D-5L and ASCOT but are difficult to directly compare given the different methodologies. A key point of difference is the use of survival as the unit of trading across different QoL states in the TTO approach which changes the cognitive nature of the task. TTO also produces direct values for particular states in comparison to DCE that produces binary choices for QoL states aggregated at the overall level to estimate values.

5.6.3. Study limitations and further research

This work has a number of limitations and associated opportunities for further work. Firstly, as this was exploratory work to understand respondent trading across dimensions, no form of anchoring to the latent scale values, for example by including duration [48, 50] was included. Therefore, the estimates are not anchored onto the full health – dead utility scale. Incorporating duration would be a natural extension for further work in this area to allow the values produced to be used as inputs for the estimation of QALYs. Secondly, as with all online studies, it is difficult to assess respondent engagement with completing the DCE task. To support completion, overlap across five dimensions was introduced, and shading was used. Both strategies have been shown to enhance respondent completion rates in previous studies [43,228]. Detailed qualitative information about whether imposing overlap and shading supported respondent completion is not available, but the general ordering and interpretability of the results, and the responses to the self-reported difficulty questions in the pilot launch suggest that the format used was acceptable to respondents.

This study is able to draw inferences about the relationship between HRQoL and SCRQoL as

measured by the EQ-5D-5L and ASCOT. Arguably, the methods applied in this study could reasonably be extended to forming other joint valuation indices that incorporate other conceptualisations of QoL. For example, further work could apply the methods used to assess the relationship between HRQoL and capabilities (as measured by the ICECAP-A [244]), or wellbeing. There are also other measures of HRQoL, that include both different and overlapping dimensions from those used in our research (for example the SF-6D [79, 80] describes HRQoL differently from the EQ-5D-5L [78]). Thus, our work provides evidence that the DCE valuation approach can be used to estimate joint indices of QoL. This provides significant scope to expand on such indices to produce QALY valuations that are more sensitive to the impacts of care services across a wide range of patient groups and settings. It could also be hypothesised that there are cross-cultural differences in preferences for different aspects of QoL depending on a number of factors such as the characteristics of the healthcare system, and different attitudes towards health and social care. Further research could adapt this work in different countries.

There are other approaches that could use DCE as the valuation method to understand how samples may trade-off across HRQoL and SCRQoL profiles. For example, trading within one perspective could be couched in terms of a particular profiles from the other perspective to understand how respondents perceive and respond to attributes when faced with a particular health profile or social care situation. It could be hypothesised that preferences for particular attributes of health would differ when placed in the context of different social care situations (or vice versa). This approach holds promise for future work to understand the relationship between perspectives, but was not done here as the aim of this work was to assess trading within an overall descriptive system to provide evidence to inform the potential development of a value set to improve the information available for decision making. Other valuation tasks such as BWS [142] may provide preference information from a different perspective about how populations trade between different health and non-health areas.

5.6.4. Conclusions

In conclusion, an established valuation methodology (DCE) has been used to demonstrate that it is possible to value concepts measuring different (and in some cases overlapping) aspects of QoL (health and social care related) on the same underlying scale. It has also been demonstrated that respondents trade between the different concepts. This has implications for decision making around the funding and use of interventions that have a wider impact on QoL than just that captured by a more narrowly focused health-related QALY metric. In the next chapter, a key

area of application of DCE methods for health state valuation is investigated. This is the construction of the designed experiment and establishing how the design impacts on the value produced helps extend knowledge about the methods that can be used to value health and wider QoL states within a DCE framework.

6. Comparing DCE designs that could be used to value measures of QoL

6.1. *Summary*

Chapter 5 demonstrated how DCEs can be used to value different QoL outcomes within the same valuation framework. A key feature of the development of DCEs is the design construction method used to select choice sets for inclusion in the study. There are a number of design construction methods available, and methodological decisions are required during the design construction process. The extent to which the design construction method chosen is influential in the characteristics of the value sets produced is unclear.

In health state valuation research, the impact of design construction has not been widely investigated, and there has been no systematic assessment of the design related and methodological choices made. The design is a fundamental building block and so it is important to systematically understand how design decisions affect choices. The study reported in this chapter investigates a number of DCE design construction strategies using the EQ-5D-5L as the descriptive system valued. The results of this work can be used to inform the design of further valuation studies of QoL outcomes. The study has direct relevance to the key questions of this thesis, as it demonstrate the advantages and disadvantages of different DCE design approaches that could be used to derive value sets for any future PBM developed to assess QoL. It uses an existing PBM about which a lot is understood (EQ-5D-5L) to allow for the implications of varying the design criteria to be tested.

6.2. Introduction

The descriptive systems of PBMs describe many combinations of dimensions and levels that could be compared by respondents within a DCE framework. They typically include 5-12 dimensions and between 3 and 6 levels for each dimension, resulting in a large number of health states that can be described. For example, the EQ-5D-5L [78] includes 3,125 possible health states which means that 4.88 million (3125 x 3124 / 2) choice sets consisting of two profiles are possible. In a DCE, decisions need to be made about which health states will be presented and then how they will be combined with other health states in choice sets. Therefore, selecting the subset of choice sets to be administered to respondents (or constructing the designed experiment) is a key part of the study design process to support the accurate elicitation of preferences and the subsequent estimation of values.

Constructing the designed experiment is a key step in the conduct of any DCE. This is demonstrated by flowcharts describing the DCE development process presented in Chapter 2 (**Figure 9** and **Figure 10**). Both flowcharts place the experimental design as the central stage that occurs after the development of the attributes and levels, and the choice set format, and is informed by the requirements of the analysis. For health state valuation purposes, the descriptive system is usually available, and the analysis aims to develop a utility value set. An experimental design with both statistical efficiency, and amenability to completion by respondents, and the ability to estimate all of the required parameters with a level of accuracy, is a prerequisite to data collection.

Given the importance of the constructed experiment, substantial research effort has gone into developing design construction methods. This has resulted in two broad classes of DCE design construction that can be described as theoretical or algorithm-based, and are summarised by Street and Viney [61]. A theoretical approach used in the construction of designs for the development of value sets is described as a generator developed design [62]. Using this construction, an orthogonal array [63] is taken as the starting point, and each row of level combinations within that array is used as the first health state in a choice set. The second option is constructed by making a systematic set of level changes given by a generator chosen. For the estimation of main effects, each level of each dimension is designed to appear as evenly as possible across the options in each choice set. Different frequencies of level combinations can also be built into the design for the estimation of more complex models (for example those estimating interactions between dimensions). Generator developed designs have been used to construct DCE designs to value a range of generic PBMs including the EQ-5D-3L [50], EQ-5D-5L [57], SF-6D [58], and also the cancer specific EORTC QLU C10D [144].

In an algorithm-based approach, a starting design of choice sets is selected, and this is sequentially changed to generate a more efficient design. This can be done by changing one profile at a time (a modified Fedorov algorithm), or by changing dimensions within a profile (a coordinate exchange algorithm [65]). The improvements in the design efficiency can be generated iteratively by assessing the error in the design, and retaining designs that improve this, or by randomly generating designs (defined as the random construction method) and retaining those with a certain level of pre-specified efficiency. Both types of iterative sequence, and approaches to improving design efficiency have been widely used to generate experimental designs to value PBMs [55, 59] (see also the structured review reported in Chapter 3).

The construction of algorithmic designs can be computationally intensive, and a further variation in the design process is the software used to implement the algorithms. In the field of health state valuation, several programs including routines available in statistical software such as Stata [66] and SAS [65] has been used alongside specifically designed DCE software such as Ngene [67], or user written commands in R [47]. However, whether there is any systematic impact of the different software implementations on the values estimated is unclear.

When designing an experiment, a decision also has to be made about whether or not to use prior values. Prior values are information about the size of the parameter levels for attributes included in the design. Prior information can be taken from a number of sources including pilot work, or other studies reported in the literature. For a number of reasons, including the lack of previous studies, and lack of pilot work, the valuation of unique attributes, or the inappropriateness of other estimates, priors are not always available. In this case, priors are referred to as non-informative, or zero, priors. When prior values are available, they can be described as informative or non-zero priors. Designs can also use both point priors (that are single values), or the Bayesian approach that uses an underlying distribution for the priors.

In theoretical design constructions, priors are not used, and this is equivalent to assuming that the prior value is zero. Therefore, such design constructions will select combinations of levels based on criteria, typically an objective function based on the information matrix. This is the situation for the generator developed approach. In algorithmic constructions, informative priors can be used, and the algorithm makes use of this information to select choice sets that optimise the objective function that is being maximised. Both point and Bayesian priors have been used in algorithm-based design approaches (See Chapter 3).

Overlap of levels is a design feature that has been proposed and used in DCE for health state valuation that reduces cognitive burden on respondents but comes at the cost of statistical efficiency. Level overlap means that a certain number of attributes within the choice sets appear at the same severity level in both profiles. This potentially makes the task easier as respondents can focus on the reduced number of dimensions that differ. This has been used in the valuation of EQ-5D-5L [228] and has been shown to produce value sets with monotonic estimates (with decreasing utility as the dimension severity increases).

6.3. Relevant past work comparing design constructions

There has been some previous empirical research comparing certain aspects of DCE design constructions. In wider DCE research, Burgess et al [245] compared generator developed, and optimal designs with and without the use of prior values. In field trials it was found that the average precision of the parameter estimates across the designs was comparable. In related work, Burgess et al [246] tested how sample size impacted the parameters estimated across generator developed designs of varying optimality, a SAS design and a design with choice sets randomly selected, and found that design efficiency becomes more important as sample size decreases. Domínguez-Torreiro [247] compared a generator developed design and an efficient algorithmic design in the field, and found that the values produced differed, but the overall goodness-of-fit was similar. Other empirical comparisons were conducted by Olsen and Meyerhoff [248] who compared four designs optimised for different criteria, and found that the level of preference heterogeneity estimated differed across the designs. Regarding DCE's for health state valuation, a design comparison found that using informative priors in a DCE with duration framework resulted in more inconsistent coefficient estimates than using non-informative priors [59]. This implies a trade-off between statistical and respondent efficiency in the choice of the priors. Studies comparing a number of designs developed using different construction methods for the purpose of valuing QoL have not been widely conducted. Therefore, further work in this area is required.

6.4. *Aims and objectives*

The aim of this study was to compare a number of DCEs for the purpose of valuing QoL, constructed using different design methods and choices, in a general Australian population sample. This aligns with overall Aim 5 of this thesis which is to compare DCE designs that could be used to value wider measurement systems. The designs selected were informed by the results of an earlier study assessing the design performance using simulated data [249].

The EQ-5D-5L was used as the descriptive system for this assessment, and is amenable to testing in this way given that the dimension descriptions are ordered, there are only five dimensions, and the majority of respondents understand the differences between the levels. Therefore, it is a useful set of descriptors to test coefficient ordering and heterogeneity.

6.5. Summary of methodological process undertaken

In this study a series of steps were undertaken to systematically compare DCE designs with different characteristics. A set of designs were constructed, and a number of indicators were tested using simulation methods. Following this, the designs were implemented in a general population sample, with respondents randomly allocated to one of the designs. The resulting data were analysed using a descriptive and modelling based approaches and compared across a number of features. In the sections below, the different designs, the simulation process, data collection, analysis and results comparison are described. The results are then discussed.

6.6. Methods – Summary of design construction features

The study uses 19 designs characterised by different design construction methods with different combinations of key features. **Table 49** describes all 19 designs each of which included 50 choice sets. The design features considered within each construction methods were the level of overlap, the use of prior information, and the implementation platform of the algorithm. Two further design properties (level balance and number of dominated choice sets) were also considered. Details about each of the features and design properties tested is provided below. Following this, the design construction methods used are described. The 19 designs developed are coded from A to S. These include seven overlap designs (A – G) and 12 non-overlap designs (H – S).

Table 49: The 19 designs included in data collection

Study design code	Overlap	Platform/method	Prior	Level balance	N dominant pairs
A	Overlap on 2	Generator	Zero Prior	0	23
B	Overlap on 2	Ngene Fedorov	Zero prior	1,870	9
C	Overlap on 2	Ngene Fedorov	Krabbe Prior	2,290	6
D	Overlap on 2	SAS Fedorov	Zero prior	1,750	11
E	Overlap on 2	SAS Fedorov	Krabbe Prior	1,730	15
F	Overlap on 2	Oppe	Zero prior	360	0
G	Overlap on 2	Oppe	Krabbe Prior	260	0
H	No overlap	Generator	Zero Prior	0	7
I	No overlap	Generator	Zero Prior	0	10
J	No overlap	Ngene Fedorov	Zero prior	210	2
K	No overlap	Ngene Fedorov	Krabbe Prior	400	0
L	No overlap	Ngene Swapping	Zero prior	0	3
M	No overlap	Ngene Swapping	Krabbe Prior	0	0
N	No overlap	Stata Fedorov	Zero prior	80	6
O	No overlap	Stata Fedorov	Krabbe Prior	80	0
P	No overlap	SAS Fedorov	Zero prior	50	3
Q	No overlap	SAS Fedorov	Krabbe Prior	190	2
R	No overlap	Oppe	Zero prior	390	0
S	No overlap	Oppe	Krabbe Prior	560	0

6.7. Methods - Design construction method

Four design construction methods were tested, including one theoretical and three algorithmic approaches. These are described in detail below.

6.7.1. Level of overlap

The level of overlap relates to the number of dimensions designed to be at the same severity level across both options in a choice set. In this study, the level of overlap was specified to be two dimensions (for seven designs; A to G), or all five dimensions were allowed to vary (for 12 designs; H to S). This was done to increase choice consistency by reducing task complexity although it does reduce the level of statistical efficiency.

6.7.2. Use of priors

Two prior values were used in this study. The first was a set of uninformative priors which were equal to zero. The other set of priors were informative, and were taken from a published DCE without duration study conducted in the UK, the Netherlands, Canada and Spain [143]. The prior values (described as the "Krabbe priors"), and standard errors around the values, are replicated in **Table 50**.

Table 50: The Krabbe priors (mean and standard error)

	MO		SC		UA		PD		AD	
	M	SE	M	SE	M	SE	M	SE	M	SE
No problems	0	0	0	0	0	0	0	0	0	0
Slight problems	-0.299	0.031	-0.208	0.033	-0.194	0.032	-0.244	0.033	-0.195	0.034
Moderate problems	-0.349	0.035	-0.290	0.035	-0.254	0.035	-0.244	0.035	-0.454	0.035
Severe problems	-0.923	0.036	-0.793	0.036	-0.769	0.035	-1.017	0.036	-1.183	0.037
Extreme problems	-1.326	0.039	-0.966	0.035	-0.987	0.035	-1.258	0.036	-1.401	0.038

MO: Mobility; SC: Self-Care; UA: Usual Activities; PD: Pain and Discomfort; AD: Anxiety and Depression

6.7.3. Implementation platform

A range of implementation platforms were used to develop the algorithmic designs. These included the specialist DCE design software Ngene [67], and macros or user written code developed in SAS [65], Stata [250] and R [47].

6.7.4. Design property assessed – Level balance

Level balance is an indicator of the number of times each level from each dimension appears in the overall design, where within each design, each dimension has 100 appearances across the

50 choice sets. Therefore, the largest value that level balance can take is 200,000, which occurs when only one of the five levels appears in each of the five attributes. When all the levels of all of the attributes appear equally often then the level balance value is zero.

6.7.5. Design property assessed - Dominated choice sets

The EQ-5D has attributes in which there is a clear ordering, from best to worst, on the levels of each of the attributes. Thus, if we have two distinct health states, but one of the health states is better or the same on all response levels, then this health state is said to dominate (and should result in all respondents choosing it as their preferred option). The number of choice sets with dominant states in each design was assessed.

6.8. Methods - Design construction methods

Four design construction methods were tested, including one theoretical and three algorithmic approaches. These are described in detail below.

6.8.1. Generator developed

Generator developed methods [62] were used for three designs (A, H and I). This theoretical approach to design started with a set of profiles and an associated set of generators. Each generator was then added in turn to the initial profile to generate choice sets of size two. For the EQ-5D-5L the initial profiles formed an orthogonal array representing the five dimensions each with five levels. This approach has been used to value a range of PBMs in Australia [50, 57].

The properties of the choice sets depend on the properties of the profiles and generators. In this study, the same set of initial profiles was used for the three generator developed designs, but the generators had different properties. As design A required overlap on two dimensions, each of the two generators had to have two entries of zero. Therefore, the generators used were (1,1,3,2,2) and (2,2,2,1,1). This resulted in 23 dominated pairs. Designs H and I are optimal for non-informative priors, and design H used the generators (1,1,1,2,2) and (2,2,2,4,4), and Design I used the generators (1,1,1,1,1) and (2,2,2,2,2). This led to seven and ten dominated pairs respectively.

6.8.2. Modified Fedorov algorithm

Modified Fedorov algorithms were used to develop ten designs in this study, four with overlap (B, C, D, E), and six with no overlap (J, K, P, Q, N, O). This algorithm is an adaptation by Zwerina

et al. [251] of the modified Fedorov [230] algorithm developed by Cook and Nachtsheim [229]. Following this approach, a set of choice sets were selected from the full factorial (non-overlap designs), or a partial factorial including all possible options with two dimensions of the five at the same level (overlap designs). Then the profiles in the design were sequentially replaced by other options from the candidate set, and any exchange that improves the efficiency of the design was retained. A full iteration was completed when all of the profiles in the choice sets included in the design have been considered. The algorithm was continued until no substantial improvement was observed in each iteration.

6.8.3. Coordinate exchange algorithm

The coordinate exchange algorithm was used for two designs without overlap in this study (L and M). This method started with a random collection of choice sets, and then changed features of the attributes and levels for the whole design by swapping the levels with other possible alternatives. The swapping algorithm was introduced by Huber and Zwerina [64]. Due to the swapping process, this approach only allows for designs without overlap to be developed. Swapping algorithms have been used in the previous valuation of PBMs using DCE method [59, 151].

6.8.4. Random selection

Random selection was used for two overlapping (F and G) and two non-overlapping (R and S) designs. These designs were developed by study collaborator Dr Mark Oppe, who provided them for inclusion in the study. The random selection algorithm was used to construct the design for the international protocol for the valuation of the EQ-5D-5L [47]. The designs were generated randomly, with the constraints that no pair could be repeated (order of options within the choice sets is not relevant), no choice sets could contain a dominated option and the designs had to satisfy a pre-specified criteria for level balance. Designs were constructed assuming a normal prior distribution centred at either 0, for the non-informative, or at the assumed informative prior value, with the SD assumed to be 0 (non-informative) or linked to the informative prior values used. The number of designs tested depended on both the assumed prior and level of overlap. For each design, the D-Error was calculated for a maximum of 20,000 (this number was arbitrarily chosen) designs. The D-Error is calculated by scaling the determinant of the variance covariance matrix by the number of parameters in the model. In total about 970,000 designs were created for each of the designs with no overlap (and about 950,000 were rejected due to level balance being greater than 900) and about 5,120,000 (actual numbers 5,129,488 and

5,125,193) for each of the designs with overlap on two attributes (of which about 5,100,000 were rejected due to level balance being greater than 900).

It should be noted that the designs specified to not have any overlap (R and S) did actually include choice sets with different levels of overlap. For Design R, 20 choice sets overlapped on one attribute, three overlapped on two attributes, and one overlapped on three. For Design S, 24 overlapped on one attribute, six on two attributes and one on three attributes.

6.9. Design combinations excluded

Due to the nature of the design process, or platform differences, a number of combinations could not be used to generate a design and were therefore excluded. This includes using a generator developed design with informative priors, and constructing an overlap design using an Ngene swapping algorithm or the software Stata.

6.10. Summary of the simulation process

The 19 designs tested in this study were also examined using simulation methods to ensure that they were amenable to the production of ordered and feasible coefficient decrements, and were able to recover assumed prior values. The simulation code was developed by Prof Deborah Street, a supervisor of this PhD. In the simulations conducted, data were generated assuming an MNL model using both the informative or non-informative priors used to construct the designs, and with extreme value type 1 errors. The deterministic component of the utility function was used to calculate the probability weights associated with each of the responses within each choice set.

The simulation process generated 1,000 sets of 3,750 responses for each prior. The number of responses was selected to represent the approximate number of observations expected for each of the 19 designs (see Section 6.12). Generating 1,000 sets of simulated betas is in line with recent work in the area [252,253].

Various indicators were generated from the simulated data to provide an indication of design performance. These were adapted from previous simulation studies [254, 255], and included the standardised bias, the root mean square error (RMSE), and the coverage. Between them these measures provide some information regarding the ability of the design to recover the

parameters assumed in the data generation process, regardless of the priors assumed when generating the design.

Table 51: Simulation indicators for the 19 designs

	Zero prior			Krabbe prior		
Design	**Bias[a]**	**RMSE[b]**	**Coverage**	**Bias**	**RMSE**	**Coverage**
A	0.027	0.096	0.947	0.055	0.108	0.950
B	0.021	0.092	0.951	0.060	0.107	0.951
C	0.032	0.096	0.952	0.029	0.108	0.946
D	0.032	0.085	0.951	0.042	0.101	0.952
E	0.022	0.161	0.953	0.043	0.186	0.948
F	0.027	0.105	0.951	0.062	0.119	0.950
G	0.027	0.099	0.953	0.048	0.112	0.953
H	0.042	0.065	0.950	0.070	0.085	0.946
I	0.037	0.065	0.951	0.135	0.093	0.954
J	0.020	0.067	0.952	0.067	0.085	0.949
K	0.025	0.067	0.948	0.049	0.085	0.951
L	0.020	0.066	0.952	0.053	0.086	0.954
M	0.039	0.082	0.954	0.046	0.086	0.949
N	0.022	0.066	0.951	0.064	0.083	0.949
O	0.021	0.085	0.950	0.048	0.088	0.946
P	0.025	0.066	0.951	0.057	0.085	0.954
Q	0.026	0.150	0.953	0.039	0.182	0.953
R	0.025	0.085	0.951	0.055	0.106	0.946
S	0.017	0.085	0.948	0.076	0.111	0.949

[a] Standardised Bias; [b] Root Mean Squared Error

6.10.1. Standardised bias

The standardised bias assesses the deviation of an estimate (in this case the estimate of the prior parameter from the simulation) from the actual (in this case the assumed beta from the prior value), and is calculated as in Equation 14:

$$\overline{\widehat{\boldsymbol{\beta}}_{\boldsymbol{\iota}}} = \boldsymbol{\Sigma}_{\boldsymbol{q}}\, \widehat{\boldsymbol{\beta}}_{\boldsymbol{iq}} \boldsymbol{/\ 1000} \quad (14)$$

Where $\hat{\beta}_{iq}$ is the estimate for β_i where i takes a value between 1 and 20 (corresponding to the 20 parameters estimated in each simulation), and q is a particular simulation. The standardised bias for β_i is given by Equation 15:

$$(\overline{\widehat{\boldsymbol{\beta}}_{\boldsymbol{\iota}}} - \boldsymbol{\beta_i})^2 + \boldsymbol{SE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}) \quad (15)$$

In **Table 51**, the value recorded is provided by Equation 16:

$$\overline{\boldsymbol{b}} = \boldsymbol{\Sigma_i b_i} \ / \ \boldsymbol{20} \quad (16)$$

6.10.2. Root Mean Squared Error

The RMSE provides an indicator of the overall accuracy as it includes both bias and variability which is transformed back onto the same scale as the parameter estimates. The RMSE for each parameter i is given by Equation 17:

$$\boldsymbol{RMSE_i} = \sqrt{[(\overline{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}} - \boldsymbol{\beta_i})^2 + \boldsymbol{SE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})^2]} \quad (17)$$

In **Table 51** the RMSE is calculated by dividing the total RMSE total for each design by the number of parameters estimated as in Equation 18:

$$\overline{\boldsymbol{RMSE}} = \boldsymbol{\Sigma_i RMSE_i} \,/\, \boldsymbol{20} \quad (18)$$

6.10.3. Coverage

The coverage is the proportion of 95% confidence intervals that include the prior value. Ideally this value would be 0.95 and it has been suggested [254] that a criterion for an acceptable coverage is that the proportion lie within two standard deviations of the nominal proportion, that is, between 0.936 and 0.964 for this study. Under- and over-coverage impacts on the Type 1 and Type 2 error rates. The coverage was recorded as the proportion of the 20 x 1000 confidence intervals that cover the assumed beta. **Table 51** reports the simulation indicators for each design. The zero prior designs have lower standardised bias, and lower RMSE. Of the overlap designs, Designs E (SAS Fedorov with Krabbe priors) and F (random with zero priors) have a larger RMSE. All coverage results are in the acceptable range.

6.11. Methods - Study design

A parallel arm study design was used, where each respondent was randomly allocated to complete 20 choice sets from one of the designs. The 20 choice sets were also randomly allocated from the overall pool of 50. To assess response consistency, one of the choice sets was also repeated as the 21st. This was either the 10th or the 12th task completed. The survey included a number of subsections being (in order) basic demographics (age, gender and region) for quota purposes, survey information and consent, self-reported EQ-5D-5L, instructions about the DCE tasks, 21 DCE choice sets, follow-up questions (around survey difficulty and overlap) and further demographic questions.

6.12. Methods - Sample size and respondents

The target sample size was 3,000, recruited through an online panel representative of the Australian general population in age, gender and region (at the state and territory level). This sample size was selected to result in around 63 observations per choice set which is consistent

with other studies valuing PBMs using DCE methods. This was calculated by working out the total number of observations (3,000 respondents * 20 choice sets = 60,000 observations) and dividing this by the total number of choice sets (950). To ensure a generally representative sample across the 19 designs, quota allocation was done at the design level. Respondents fully completing the survey were provided with a small incentive from the panel company. This is process was approved by the Centre for Health Economics Research and Evaluation Program Ethics Process.

6.13. Methods - Data analysis

The data were analysed using a range of modelling approaches including conditional logit (which has been employed widely to model EQ-5D DCE data), and models assuming heterogeneity of preferences.

6.13.1. Comparing sample characteristics

The demographic characteristics of the sample were compared descriptively across the 19 designs. Chi Square tests were used to compare the proportions of respondents in each demographic group across the designs.

6.13.2. Assessing respondent behaviour

Three potential indicators of the impact of different designs on respondent completion behaviours were assessed descriptively. The first behaviour examined was the time taken to complete the DCE choice sets for the overall sample and for each design. Time taken may be a proxy for respondent attention, and could demonstrate the impact of design issues such as choice set difficulty on completion. The significance of time taken to complete the tasks was compared between the overlap and non-overlap designs, and the non-informative and informative priors, using one-way analysis of variance (ANOVA) difference testing. Overlap was used an indicator of possible differences in time taken given this was the design criterion that may support respondent completion by simplifying the task (and therefore impact the time taken). The type of prior information was also used to test time taken given that different priors may also impact the complexity of the design. Using priors to divide the sample also resulted in two relatively large groups (8 designs used informative priors, and 11 used non-informative priors. Other design criteria were not used to assess time taken given they were not specified to improve respondent efficiency, and the smaller number of designs specified for each criteria

would make comparisons more difficult to interpret (for example, comparing theoretical and algorithmic designs would result in a comparison of groups of 3 and 6 designs respectively).

In the DCE study reported in chapter 5, analysis of the impact of different respondent completion times on the parameter estimates were conducted. A similar analysis was not conducted in this study given that dividing the sample by time taken would reduce the respondents in each group to a point where the models would become challenging to interpret. The second indicator assessed was the extent of drop out from the survey overall and at the design level, and whether the drop out was initiated by the respondent or as a system time out. High dropout could also be indicative of task difficulty. The third indicator assessed the consistency of responses for each participant using the repeated question (where the task that appeared 10^{th} or 12^{th} was repeated as the 21^{st}).

6.13.3. Comparing designs - Assessing feedback questions

Descriptive analyses and Chi Square difference testing were used to compare respondent self-report questions by whether they completed tasks from an overlapping or non-overlapping design. The level of overlap specified for the design was used as the grouping variable given that overlap has been used to simplify choice sets for respondents in other EQ-5D-5L DCE valuation studies [44].

6.13.4. Comparing designs – Conditional logit analysis

A main effects model was estimated for each design using conditional logit regression with robust standard errors (to take into account repeat observations per person). Models were estimated for all 19 designs separately, and the consistency and significance of the coefficients within each dimension across the designs was compared. The overall dimension preference order was also investigated as a descriptive comparison of using the magnitude of the coefficient decrement at level 5 as a proxy for dimension importance. Conducting these comparisons provides information about whether a certain design feature leads to more or less non-monotonic and significant coefficients, or different overall dimension level preference patterns. Analysis was conducted using Stata V15 [66].

However, as described in Section 2.9.3, the magnitude of the DCE values such as those produced in this study are estimated on a latent scale, and therefore cannot be directly compared in terms of the overall size of the estimates. To allow for scale free comparison, the conditional logit

parameter estimates were anchored by dividing them by the estimated latent value for the worst EQ-5D-5L health state (55555) produced for each design. The value was calculated by summing the level five estimate from each of the five dimensions.

6.13.5. Comparing designs - Assessing poolability

To assess whether the scale of the designs was comparable, the pooled data for the seven overlap and 12 non-overlap designs were tested separately using the scale assessment approach outlined by Swait and Louviere [71], and described in detail in Section 2.9.11. The method was applied to test scale across a range of demographic groups in the DCE study described in Chapter 5. In this study, the focus was on testing the difference in scale across the designs, as testing by demographic group would reduce the sample size in each subgroup, and the subsequent stability of the models produced.

6.13.6. Comparing designs – Assessing preference heterogeneity using latent class

Latent class analysis was used to assess preference heterogeneity across the 19 designs. The latent class approach is described in detail in Section 2.9.12. The rationale for testing the results of latent class analysis across the different designs was to understand whether the design characteristics impacted the ability to estimate classes, and subsequently to compare the characteristics of the classes across the designs.

In this study, the latent class modelling was informed by the analysis conducted in the DCE research described in Chapter 5. Models with two to five classes were tested for each design. The BIC was favoured as the indicator of the number of classes to extract for each, in line with the DCE study reported in Chapter 5. The BIC is a model fit indicator that takes both the number of parameters and observations into account, and is described in Section 5.4.18). Five was selected as the largest class number given the expected, and subsequently observed, sample size for each design (approximately 180 per design). Demographic indicators for age, gender and condition were included in the models, and represented as probabilities of each demographic belonging to each class. The model with the lowest BIC was estimated for each design, and similarities and differences in the class structures and characteristics across the different design features were assessed.

6.13.7. Comparing designs – Assessing preference heterogeneity using mixed logit

Preference heterogeneity was also assessed across each design separately using mixed logit regression [73], which is described in Section 2.9.13. In the mixed logit models estimated, the parameters for all 20 EQ-5D-5L dimension levels estimated (excluding the baseline level) were specified to be heterogeneous. The rationale for doing this was to understand whether a certain design type was more likely to be sensitive to differing levels of preference heterogeneity (given similarities in the demographic composition of the samples completing each design). This analysis was informed by the exploratory mixed logit modelling conducted in Chapter 5 which suggested that models with a higher number of Halton Draws (1,000), and a burn rate at least as large as the number of parameters specified as random (20) were favoured. The parameters were specified to be uncorrelated given the issues with estimating correlated coefficients encountered in that study.

As part of the mixed logit analysis the predicted probabilities of respondents choosing option A in each choice set was also estimated for each design, and compared to the observed frequency. The mean and range of the probabilities for each choice set, was calculated. It was then assessed whether the observed frequency of choosing option A was within the range of the predicted probability at the overall level for each choice set within each design. This provided 19 overall scores between 0 (all observed frequencies outside predicted range) and 50 (all observed frequencies within predicted range) for each design which was then converted into a percentage to allow for comparisons across designs. The mixed logit and predictive analysis was conducted using the mixlogit command in Stata [247].

6.14. Results

6.14.1. Completion and sample characteristics

Table 52 reports the characteristics of the sample overall and for each of the seven overlap designs, and **Table 53** reports the characteristics for the 12 non-overlapping designs. The characteristics of the overall sample are similar to the overall Australian population in terms of age, gender and region at the state and territory level. The respondents are more highly educated than the overall Australian population. At the overall level, 47% of the sample report having a long-term health condition, and 22% report themselves to be in the best EQ-5D-5L health state. There are no significant differences between the subsamples at the 0.05 level by age, gender or region for the overlap or non-overlap groups. This is also the case for the majority of the other demographic characteristics measured.

6.14.2. Comparing designs – Respondent behaviour

Table 54 reports the descriptive results of the time taken to complete the survey for the overall sample and for each design. The overall time taken does not significantly differ across all designs, but the mean time taken does significantly differ ($F_{(1,3353)} = 8.24$, $p = 0.004$) between the designs with overlap (428 seconds) and non-overlap (391 seconds). The mean time taken between the designs with non-informative (399 seconds) and informative priors (414 seconds) does not significantly differ ($F_{(1,3353)} = 1.36$, $p = 0.244$). The means are longer than the medians given evidence of outliers taking a substantial amount of time to complete the survey. At the 95th percentile, the shortest time is 808 seconds (Design B), and the longest in 1,310 seconds (Design C). The median time taken ranges from 291 seconds (Design H) to 366 seconds (Design C).

Table 52: Demographic characteristics of the overall sample, and seven overlap designs

Characteristic	Overall	A	B	C	D	E	F	G	Sig[a]
N	3,363	182 (5)	177 (5)	173 (5)	181 (5)	180 (5)	177 (5)	176 (5)	NS[b]
Male	1,596 (48)	86 (47)	84 (47)	82 (47)	82 (45)	86 (48)	84 (48)	84 (48)	NS
Age category									NS
18 – 29	765 (23)	43 (24)	41 (23)	40 (23)	37 (20)	45 (25)	44 (25)	40 (23)	
30 – 39	588 (17)	32 (18)	31 (18)	29 (17)	35 (19)	30 (17)	26 (15)	33 (19)	
40 – 49	574 (17)	26 (14)	29 (16)	26 (15)	32 (18)	35 (20)	24 (14)	27 (15)	
50 – 59	534 (16)	31 (17)	30 (17)	31 (18)	29 (16)	24 (13)	35 (20)	30 (17)	
60 – 69	451 (13)	23 (13)	26 (14)	26 (15)	27 (15)	24 (13)	22 (12)	22 (13)	
70+	451 (13)	27 (15)	20 (11)	21 (12)	21 (12)	22 (12)	26 (15)	24 (14)	
Region									NS
Aust Capital Territory	58 (2)	4 (2)	3 (2)	2 (1)	3 (2)	3 (2)	1 (1)	2 (1)	
New South Wales	1,049 (31)	61 (34)	51 (29)	52 (30)	57 (32)	53 (30)	56 (32)	50 (28)	
Northern Territory	25 (1)	1 (1)	2 (1)	1 (1)	0	2 (1)	2 (1)	0	
Queensland	659 (20)	35 (19)	41 (23)	28 (16)	32 (18)	35 (19)	34 (19)	47 (27)	
South Australia	287 (9)	15 (8)	16 (9)	17 (10)	22 (12)	15 (8)	12 (7)	12 (7)	
Tasmania	94 (3)	7 (4)	5 (3)	6 (3)	4 (2)	3 (2)	4 (2)	4 (2)	
Victoria	898 (27)	48 (26)	48 (27)	48 (28)	47 (26)	49 (27)	51 (29)	47 (27)	
Western Australia	293 (9)	11 (6)	11 (6)	19 (11)	16 (9)	20 (11)	17 (10)	14 (8)	
Health questions									
Has condition	1,537 (47)	70 (41)	79 (46)	88 (52)	81 (47)	76 (45)	78 (45)	77 (45)	NS
In EQ-5D-5L 11111	737 (22)	38 (21)	43 (24)	28 (16)	46 (25)	41 (23)	32 (18)	28 (16)	NS
Health status									NS
Good – Excellent	2,526 (77)	137 (78)	130 (75)	123 (73)	131 (75)	135 (79)	136 (79)	132 (76)	
Fair - poor	743 (23)	37 (21)	42 (24)	46 (27)	44 (25)	35 (21)	37 (21)	41 (24)	
Education									
Secondary highest	1,024 (31)	59 (32)	58 (33)	54 (31)	63 (35)	57 (32)	49 (28)	56 (32)	NS
Currently studying	505 (15)	31 (17)	33 (19)	28 (17)	29 (16)	35 (20)	23 (13)	22 (13)	NS
Married	1,886 (56)	88 (48)	94 (53)	99 (57)	94 (52)	92 (51)	100 (57)	111 (63)	NS
Gross income									NS
$0 to $80K AUD	2,382 (82)	126 (81)	121 (83)	124 (800)	131 (83)	128 (84)	136 (85)	125 (82)	
$80K +	538 (18)	30 (19)	25 (17)	31 (20)	27 (17)	24 (16)	25 (16)	28 (18)	

[a] Significance; [b] non-significant (significance greater than 0.1)

Table 53: Demographic characteristics of the overall sample, and the 12 non-overlap designs

Characteristic	Overall	H	I	J	K	L	M	N	O	P	Q	R	S	Sig[a]
N	3,363	180 (5)	175 (5)	178 (5)	171 (5)	177 (5)	173 (5)	175 (5)	177 (5)	175 (5)	178 (5)	181 (5)	177 (5)	NS[b]
Male	1,596 (48)	88 (50)	84 (48)	85 (48)	82 (48)	83 (47)	82 (47)	80 (47)	83 (47)	83 (47)	85 (48)	86 (48)	85 (48)	NS
Age category														NS
18 – 29	765 (23)	39 (22)	38 (22)	41 (23)	37 (22)	36 (20)	35 (20)	44 (25)	43 (24)	38 (22)	43 (24)	38 (21)	43 (24)	
30 – 39	588 (17)	32 (18)	32 (18)	29 (16)	32 (19)	35 (20)	33 (19)	28 (16)	28 (16)	31 (18)	29 (16)	34 (19)	29 (16)	
40 – 49	574 (17)	33 (18)	32 (18)	34 (19)	28 (16)	41 (23)	26 (15)	29 (17)	31 (18)	35 (20)	34 (19)	25 (14)	27 (15)	
50 – 59	534 (16)	27 (15)	26 (15)	26 (1)	29 (17.0)	18 (10)	32 (19)	28 (16)	25 (14)	23 (13)	24 (14)	34 (19)	32 (18)	
60 – 74	451 (13)	23 (13)	22 (13)	25 (14)	22 (13)	22 (12)	24 (14)	27 (15)	19 (11)	24 (14)	25 (14)	28 (16)	20 (11)	
75+	451 (13)	26 (14)	25 (14)	23 (13)	23 (14)	25 (14)	23 (13)	19 (11)	31 (18)	24 (14)	23 (13)	22 (12)	26 (15)	
Region														NS
Aust Capital Territory	58 (2)	3 (2)	6 (3)	3 (2)	3 (2)	1 (1)	2 (1)	6 (3)	3 (2)	5 (3)	3 (2)	3 (2)	2 (1)	
New South Wales	1,049 (31)	55 (31)	58 (33)	61 (34)	61 (36)	57 (32)	51 (29)	51 (29)	53 (30)	53 (30)	52 (29)	54 (30)	63 (36)	
Northern Territory	25 (1)	3 (2)	0	2 (1)	0	1 (1)	1 (1)	3 (2)	1 (1)	1 (1)	2 (1)	1 (1)	2 (1)	
Queensland	659 (20)	32 (18)	32 (18)	37 (21)	30 (18)	33 (19)	33 (19)	33 (19)	39 (22)	31 (18)	36 (20)	38 (21)	33 (19)	
South Australia	287 (9)	12 (7)	15 (9)	16 (9)	17 (10)	19 (11)	19 (11)	14 (8)	10 (6)	12 (7)	20 (11)	11 (6)	13 (7)	
Tasmania	94 (3)	9 (5)	3 (2)	2 (1)	8 (5)	4 (2)	5 (3)	5 (3)	5 (3)	6 (3)	5 (3)	4 (2)	5 (3)	
Victoria	898 (27)	53 (30)	52 (30)	41 (23)	40 (23)	50 (28)	40 (23)	47 (27)	50 (28)	48 (27)	42 (24)	51 (28)	46 (26)	
Western Australia	293 (9)	13 (7)	9 (5)	16 (9)	12 (7)	12 (7)	22 (13)	16 (9)	16 (9)	19 (11)	18 (10)	19 (11)	13 (7)	
Health questions														
Has condition	1,537 (47)	83 (47)	78 (45)	79 (46)	88 (53)	81 (47)	85 (50)	83 (49)	94 (54)	85 (49)	77 (45)	85 (49)	70 (41)	NS
In EQ-5D-5L 11111	737 (22)	38 (21)	39 (22)	45 (25)	41 (24)	40 (23)	33 (19)	44 (25)	29 (16)	37 (21)	48 (27)	41 (23)	46 (26)	NS
General health														NS
Good – excellent	2,526 (77)	130 (75)	131 (76)	128 (74)	132 (79)	138 (81)	124 (73)	138 (81)	130 (75)	141 (83)	135 (79)	136 (78)	138 (80)	
Fair - poor	743 (23)	44 (25)	42 (24)	45 (26)	36 (21)	33 (19)	46 (27)	33 (19)	44 (25)	30 (18)	36 (21)	38 (22)	34 (20)	
Education														
Secondary highest	1,024 (30)	51 (28)	58 (33)	42 (24)	49 (29)	47 (27)	52 (30)	61 (35)	44 (25)	52 (30)	64 (36)	56 (31)	52 (30)	NS
Currently studying	505 (15)	26 (15)	28 (16)	27 (15)	22 (13)	21 (12)	29 (17)	22 (13)	23 (13)	27 (16)	28 (16)	26 (15)	25 (15)	NS
Married	1,886 (56)	101 (56)	84 (48)	107 (60)	95 (56)	100 (57)	106 (63)	93 (53)	101 (57)	103 (59)	110 (62)	95 (53)	110 (62)	NS
Gross income														NS
$0 to $80K AUD	2,382 (82)	128 (85)	123 (80)	133 (83)	128 (85)	123 (82)	129 (82)	124 (85)	120 (78)	120 (78)	126 (82)	120 (80)	113 (74)	
$80K +	538 (18)	23 (15)	30 (20)	27 (17)	24 (19)	28 (19)	28 (18)	22 (15)	34 (22)	34 (22)	28 (18)	31 (21)	39 (26)	

[a] significance; [b] non-significant (significance greater than 0.1)

Table 54: Time taken overall and by design (in seconds)

Design	Mean (SD)	Median	5th percentile	95th percentile
Overall	405 (365)	320	90	1,007
A	457 (453)	342	82	1,405
B	380 (321)	318	90	808
C	463 (414)	366	74	1,310
D	401 (371)	321	96	857
E	449 (465)	346	121	1,193
F	425 (401)	341	81	918
G	423 (410)	337	94	1,042.6
H	362 (332)	291	86	1,059
I	404 (381)	311	79	932
J	366 (288)	303	94	902
K	356 (250)	297	90	857
L	398 (339)	320	96	1,070
M	406 (332)	323	96	1,046
N	413 (426)	321	82	1,033
O	439 (364)	336	86	1,169
P	409 (419)	316	95	1,140
Q	389 (330)	293	75	1,105
R	368 (256)	315	100	764
S	381 (276)	312	85	1,007
Overlap	428 (408)	335	92	994
Non-overlap	391 (337)	312	90	1,007
Non-informative priors	398 (367)	318	90	949
Informative priors	413 (363)	324	90	1,042

Table 55: Drop out overall and by design (in seconds)

Design	Total started survey	Comp (%)	User initiated (%)	System initiated (%)
Overall	3833	3363 (87.7)	40 (1.0)	430 (11.2)
A	207	182 (87.9)	4 (1.9)	21 (10.1)
B	209	177 (84.7)	3 (1.4)	29 (13.9)
C	201	173 (86.1)	2 (1.0)	26 (12.9)
D	199	181 (91.0)	1 (0.5)	17 (8.5)
E	205	180 (87.8)	3 (1.5)	23 (10.7)
F	197	177 (89.9)	3 (1.5)	17 (8.6)
G	194	176 (90.7)	3 (1.5)	15 (7.7)
H	206	180 (87.4)	1 (0.5)	25 (12.1)
I	203	175 (86.2)	1 (0.5)	27 (13.3)
J	199	178 (89.4)	2 (1.0)	19 (9.6)
K	193	171 (88.6)	0 (0)	22 (11.4)
L	206	177 (85.9)	5 (2.4)	24 (11.7)
M	198	173 (87.4)	2 (1.0)	23 (11.6)
N	203	175 (86.2)	2 (1.0)	26 (12.8)
O	197	177 (89.9)	3 (1.5)	17 (8.6)
P	201	175 (87.1)	0 (0)	26 (12.9)
Q	210	178 (84.8)	2 (1.0)	30 (14.3)
R	196	181 (92.4)	1 (0.5)	14 (7.1)
S	209	177 (84.7)	2 (1.0)	30 (14.4)

Table 55 reports the drop out overall and by design. The completion rates (of those starting the survey) are high, ranging from 84.7% (Design B) to 92.4% (Design R). The level of drop out does not significantly differ across the designs at the overall or across the overlap and non-overlap groups. The majority of dropouts were 'system initiated' meaning that the survey is closed due to timing out. Very few dropouts were user initiated. These results suggest that the design features explored do not influence drop out.

The percentage of respondents failing the consistency assessment by providing a different answer for the repeated task ranged from 14.4 (Design E) to 29.5 (Design M). A recent synthesis of 16 DCE studies using a repeated consistency task found a mean (SD) rate of 30% (26%) providing a different answer to the repeated question[256]. Therefore, the inconsistency rate in this study is generally lower than that observed in other work, but this comparison should be considered in light of potential methodological differences between the studies. It is worth noting that due to survey coding issues, five design surveys (Design A to E) had the 10^{th} task repeated, and 13 designs had the 12^{th} task repeated. One design (N) did not include a repeat choice set. The mean proportion of respondents failing the consistency task if it was the 10^{th} or the 12^{th} choice set (21.8% vs. 22%) did not differ.

6.14.3. Comparing designs – Feedback questions

Table 56 displays the results of the Chi Square analysis comparing the feedback question findings across the overlap level groups. Those who completed an overlap design were significantly more likely to disagree that the tasks were difficult, and also disagree that the health states were difficult to imagine. There was no difference between the groups for the other two questions (difficult to tell the difference between the options, and that they considered the whole description). These results may indicate that to some extent, respondents found the overlapping tasks easier to complete.

Table 56: Summary of feedback questions

Feedback question	Overlap (n,%)	Non-overlap (n,%)	Sig
Task difficult			0.038
Disagree – strongly disagree	819 (68.7)	1,327 (65.1)	
Neutral – Strongly agree	373 (31.3)	710 (34.9)	
Difficult to tell difference			NS
Disagree – strongly disagree	775 (65.1)	1,262 (62.0)	
Neutral – Strongly agree	416 (34.9)	775 (38.1)	
Difficult to imagine			0.002
Disagree – strongly disagree	721 (60.5)	1,117 (54.9)	
Neutral – Strongly agree	471 (39.5)	916 (45.1)	
Consider whole description			NS
Disagree – strongly disagree	126 (10.6)	231 (11.3)	
Neutral – Strongly agree	1,066 (89.4)	1,806 (88.7)	

6.14.4. Comparing designs – Conditional logit models

Table 57, 58 and **59** report the conditional logit models for the overlap and non-overlap designs respectively, with information about the number of inconsistent and non-significant coefficients presented. The designs with the highest level of inconsistent coefficients, were developed using the SAS based and Oppe algorithmic approaches. These designs also had the highest number of non-significant coefficients, both in comparison to the baseline level one and the adjacent severity level. The generator developed designs had the lowest number of inconsistent and non-significant coefficients. The standard errors for the overlap designs are substantially larger than those with no overlap reflecting the lower efficiency of the designs.

Figure 29 and **Figure 30** display the scaled parameters for the overlap and non-overlap designs respectively. The magnitudes of the decrements are reasonably similar, and the inconsistencies, which mainly appear between levels 1 and 2, and 4 and 5, can be observed. **Figure 31** and **Figure 32** show the magnitude of the scaled level 5 coefficient for each dimension across each design (as a proxy of importance). The estimated overall importance of the dimensions does differ between design types, but the difference is not consistent between design features.

Table 57: Conditional logit comparison of seven designs with overlap

Parameter	Model 30: Design A		Model 31: Design B		Model 32: Design C		Model 33: Design D		Model 34: Design E		Model 35: Design F		Model 36: Design G	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.233***	0.105	-0.557***	0.097	-0.151	0.124	-0.467***	0.103	-0.268*	0.127	-0.429***	0.105	*0.034*	0.112
MO3	-0.579***	0.131	-0.564***	0.108	-0.446***	0.096	-0.833***	0.110	-0.854***	0.229	*-0.305***	0.101	-0.176	0.103
MO4	-1.156***	0.135	-1.142***	0.120	-0.998***	0.127	-1.185***	0.106	-1.451***	0.336	-0.935***	0.112	-1.230***	0.131
MO5	-1.604***	0.117	-1.706***	0.120	-1.324***	0.140	-1.959***	0.125	-1.716***	0.146	-1.197***	0.112	-1.172***	0.106
SC2	-0.060	0.106	-0.004	0.101	-0.100	0.105	-0.310**	0.114	-0.307**	0.115	-0.104	0.112	-0.689***	0.132
SC3	-0.520***	0.131	-0.102	0.111	-0.157	0.108	*-0.289***	0.106	-0.420*	0.208	-0.123	0.127	*-0.342****	0.106
SC4	-1.101***	0.133	-0.515***	0.109	-1.030***	0.118	-1.037***	0.115	-1.407***	0.273	-0.540***	0.131	-1.173***	0.106
SC5	-1.548***	0.115	-0.929***	0.106	-1.063***	0.110	-1.368***	0.117	-1.477***	0.132	-0.988***	0.128	*-1.119****	0.140
UA2	-0.276***	0.076	*0.073*	0.094	*0.199*	0.106	*0.155*	0.115	-0.306**	0.116	-0.146	0.120	*-0.140*	0.138
UA3	-0.513***	0.076	-0.385***	0.103	-0.442***	0.107	-0.238*	0.108	-0.461*	0.180	-0.171	0.117	-0.306**	0.110
UA4	-1.133***	0.081	-0.554***	0.095	-0.462***	0.102	-0.617***	0.109	-0.981***	0.259	-0.497***	0.124	-0.757***	0.107
UA5	-1.420***	0.081	-0.854***	0.098	-0.867***	0.118	-1.080***	0.104	-1.282***	0.123	-0.749***	0.154	-0.989***	0.107
PD2	-0.279**	0.104	-0.235*	0.112	-0.127	0.113	-0.021	0.103	-0.150	0.114	-0.142	0.114	*0.001*	0.104
PD3	-0.366**	0.133	-0.241*	0.102	-0.294*	0.118	-0.587***	0.103	-0.327	0.248	-0.391***	0.116	-0.076	0.132
PD4	-1.152***	0.139	-0.793***	0.113	-0.871***	0.120	-1.212***	0.104	-1.492***	0.329	-1.073***	0.102	-1.041***	0.129
PD5	-1.493***	0.123	-1.123***	0.094	-1.178***	0.113	-1.340***	0.113	*-1.486****	0.123	*-1.059****	0.133	*-0.964****	0.114
AD2	-0.197	0.114	-0.159	0.132	*0.045*	0.110	-0.014	0.100	-0.167	0.139	*0.129*	0.148	-0.184	0.123
AD3	-0.502***	0.134	-0.577***	0.129	-0.463***	0.109	-0.344***	0.096	-0.601*	0.285	-0.373**	0.125	-0.510***	0.105
AD4	-1.020***	0.136	-0.711***	0.128	-1.129***	0.118	-1.059***	0.104	-1.569***	0.381	-1.300***	0.133	-1.582***	0.121
AD5	-1.248***	0.121	-1.011***	0.107	-1.261***	0.112	-1.235***	0.112	*1.271****	0.115	*-1.286****	0.121	*-1.529****	0.120
No Obs[c]	7,280		7,080		6,920		7,240		7,200		7,080		7,040	
LL[d]	-2,040		-2,013		-1,996		-1,855		-1,794		-2,137		-2,000	
AIC[e]	4,120		4,067		4,027		3,750		3,628		4,315		4,040	
BIC[f]	4,258		4,204		4,164		3,888		3,765		4,452		4,177	
No incons[g]	0		1		2		2		2		4		7	
No non-sig[f]	2		4		6		3		3		6		6	

[a] Coefficient estimate; [b] standard error; [c] Number of observations; [d] Log-likelihood; [e] Akaike Information Criterion; [f] Bayesian Information Criterion; [g] number of inconsistencies; [f] number of non-significant estimates; p-values for the difference between the coefficient and baseline indicated by stars: *** 0.001, ** 0.01; *0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; italics: inconsistent coefficients

Table 58: Conditional logit comparison of 12 non-overlapping designs (part 1)

Parameter	Model 37: Design H		Model 38: Design I		Model 39: Design J		Model 40: Design K		Model 41: Design L		Model 42: Design M	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.292***	0.076	-0.366***	0.074	-0.095	0.076	-0.114	0.076	-0.173**	0.077	-0.125	0.074
MO3	*-0.281****	0.078	-0.494***	0.075	-0.279***	0.080	-0.319***	0.079	-0.500***	0.080	-0.387***	0.073
MO4	-0.728***	0.080	-0.810***	0.078	-0.847***	0.084	-0.728***	0.100	-0.918***	0.081	-0.856***	0.096
MO5	-1.107***	0.081	-1.074***	0.081	-1.226***	0.091	-1.244***	0.128	-1.313***	0.084	-1.225***	0.121
SC2	-0.167*	0.076	-0.204**	0.074	-0.292***	0.076	-0.110	0.078	*0.046*	0.078	-0.127	0.071
SC3	-0.201**	0.079	-0.332***	0.075	-0.348***	0.078	-0.210**	0.074	-0.070	0.078	-0.253***	0.073
SC4	-0.558***	0.078	-0.469***	0.077	-0.922***	0.082	-0.638***	0.097	-0.434***	0.083	-0.678***	0.091
SC5	-0.793***	0.082	-0.629***	0.080	-0.922***	0.085	-0.910***	0.104	-0.660***	0.082	-1.017***	0.100
UA2	-0.153*	0.076	-0.126	0.074	*0.059*	0.081	-0.051	0.067	*0.002*	0.079	-0.031	0.072
UA3	-0.202**	0.080	-0.217**	0.075	-0.108	0.085	-0.245***	0.073	-0.148	0.081	-0.134	0.075
UA4	-0.324***	0.080	-0.358***	0.078	-0.395***	0.078	-0.400***	0.092	-0.643***	0.081	-0.590***	0.089
UA5	-0.551***	0.081	-0.431***	0.081	-0.574***	0.075	-0.593***	0.110	-0.689***	0.083	-0.651***	0.202
PD2	-0.149*	0.073	-0.205**	0.074	-0.183*	0.081	-0.143*	0.070	-0.197*	0.080	-0.030	0.070
PD3	-0.343***	0.075	-0.349***	0.075	-0.404***	0.082	*-0.033*	0.076	-0.214*	0.085	-0.198**	0.071
PD4	-0.669***	0.074	-0.554***	0.077	-0.802***	0.082	-0.568***	0.107	-0.558***	0.078	-0.878***	0.099
PD5	-0.941***	0.079	-0.697***	0.080	-0.964***	0.086	-0.840***	0.119	-0.810***	0.082	-1.118***	0.116
AD2	-0.013	0.073	-0.112	0.074	-0.199*	0.077	-0.114	0.069	-0.139	0.079	-0.123	0.069
AD3	-0.269***	0.074	-0.339***	0.075	-0.213**	0.080	-0.383***	0.080	-0.380***	0.081	-0.428***	0.078
AD4	-0.847***	0.075	-0.550***	0.077	-0.874***	0.086	-1.023***	0.118	-0.884***	0.086	-0.997***	0.113
AD5	-0.896***	0.079	-0.708***	0.080	-1.053***	0.082	*-0.986****	0.123	-1.010***	0.086	-1.132***	0.123
No Obs[c]	7200		7000		7120		6840		7080		6920	
LL[d]	-2115		-2213		-1939		-2282		-1979		-2305	
AIC[e]	4,270		4,467		3,919		4,605		3,999		4,650	
BIC[f]	4,407		4,604		4,057		4,742		4,136		4,787	
No incons[g]	1		0		1		2		2		0	
No non-sig[f]	1		1		3		5		5		6	

[a] Coefficient estimate; [b] standard error; [c] Number of observations; [d] Log-likelihood; [e] Akaike Information Criterion; [f] Bayesian Information Criterion; [g] number of inconsistencies; [f] number of non-significant estimates; p-values for the difference between the coefficient and baseline indicated by stars: *** 0.001, ** 0.01; *0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; italics: inconsistent coefficients

Table 59: Conditional logit comparison of 12 non-overlapping designs (part 2)

Parameter	Model 43: Design N		Model 44: Design O		Model 45: Design P		Model 46: Design Q		Model 47: Design R		Model 48: Design S	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.161**	0.076	-0.350***	0.074	-0.203**	0.077	-0.231**	0.083	-0.022	0.105	-0.142	0.106
MO3	-0.387***	0.076	-0.452***	0.074	-0.279***	0.079	-0.324	0.200	-0.159	0.085	-0.199	0.105
MO4	-0.856***	0.078	-1.025***	0.102	-0.862***	0.083	-0.941***	0.318	-0.550***	0.098	-0.820***	0.108
MO5	-1.123***	0.081	-1.529***	0.130	-1.199***	0.082	-0.997***	0.113	-0.828***	0.099	-1.032***	0.111
SC2	-0.127***	0.075	*0.008*	0.072	-0.185**	0.079	-0.035	0.086	-0.100	0.095	-0.372***	0.101
SC3	-0.253***	0.077	-0.237***	0.072	*-0.184***	0.078	-0.133	0.181	*-0.075*	0.092	*0.147*	0.123
SC4	-0.677***	0.082	-0.695***	0.093	-0.675***	0.079	-0.623**	0.253	-0.548***	0.092	-0.509***	0.096
SC5	-1.017***	0.077	-0.895***	0.100	-0.936***	0.080	*-0.515****	0.095	-0.903***	0.109	-0.627***	0.111
UA2	-0.031	0.073	-0.088	0.069	-0.126	0.078	-0.005	0.080	-0.155	0.100	-0.178	0.118
UA3	-0.134	0.077	-0.201**	0.074	-0.207*	0.078	-0.478**	0.174	-0.187*	0.099	*-0.115*	0.100
UA4	-0.591***	0.078	-0.630***	0.092	-0.368***	0.080	-0.630**	0.249	-0.577***	0.090	-0.320***	0.104
UA5	-0.651***	0.079	-0.765***	0.102	-0.508***	0.084	*-0.477****	0.095	-0.833***	0.097	-0.585***	0.094
PD2	-0.030	0.077	-0.119	0.070	-0.107	0.078	-0.133	0.080	*0.042*	0.103	-0.528***	0.114
PD3	-0.198	0.079	-0.228***	0.070	-0.145	0.078	-0.328	0.221	-0.043	0.116	*-0.289****	0.093
PD4	-0.878***	0.075	-0.830***	0.106	-0.608***	0.075	-0.707*	0.310	-0.518***	0.085	-0.909***	0.107
PD5	-1.117***	0.079	-1.190***	0.121	-0.743***	0.079	-0.730***	0.100	-0.747***	0.107	-1.024***	0.085
AD2	-0.123	0.071	-0.180*	0.069	-0.384***	0.075	-0.143	0.107	-0.065	0.097	-0.157	0.097
AD3	-0.428***	0.073	-0.362***	0.074	-0.443***	0.077	-0.329	0.277	-0.433***	0.113	-0.285**	0.110
AD4	-0.997***	0.081	-1.212***	0.112	-0.943***	0.085	-0.928**	0.371	-0.998***	0.108	-0.958***	0.101
AD5	-1.132***	0.079	-1.338***	0.130	-0.968***	0.082	*-0.803****	0.803	-1.157***	0.102	*-0.880****	0.097
No Obs[c]	7350		7080		7000		7120		7240		7080	
LL[d]	-2150		-2355		-2003		-2065		-2076		-1992	
AIC[e]	4,126		4,751		4,047		4,170		4,192		4,025	
BIC[f]	4,263		4,889		4,185		4,308		4,329		4,162	
No incons[g]	0		1		1		3		2		4	
No non-sig[f]	5		3		3		8		9		6	

[a] Coefficient estimate; [b] standard error; [c] Number of observations; [d] Log-likelihood; [e] Akaike Information Criterion; [f] Bayesian Information Criterion; [g] number of inconsistencies; [f] number of non-significant estimates; p-values for the difference between the coefficient and baseline indicated by stars: *** 0.001, ** 0.01; *0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; italics: inconsistent coefficients

Figure 29: Parameter estimates scaled on the value of health state 55555 (overlap designs)

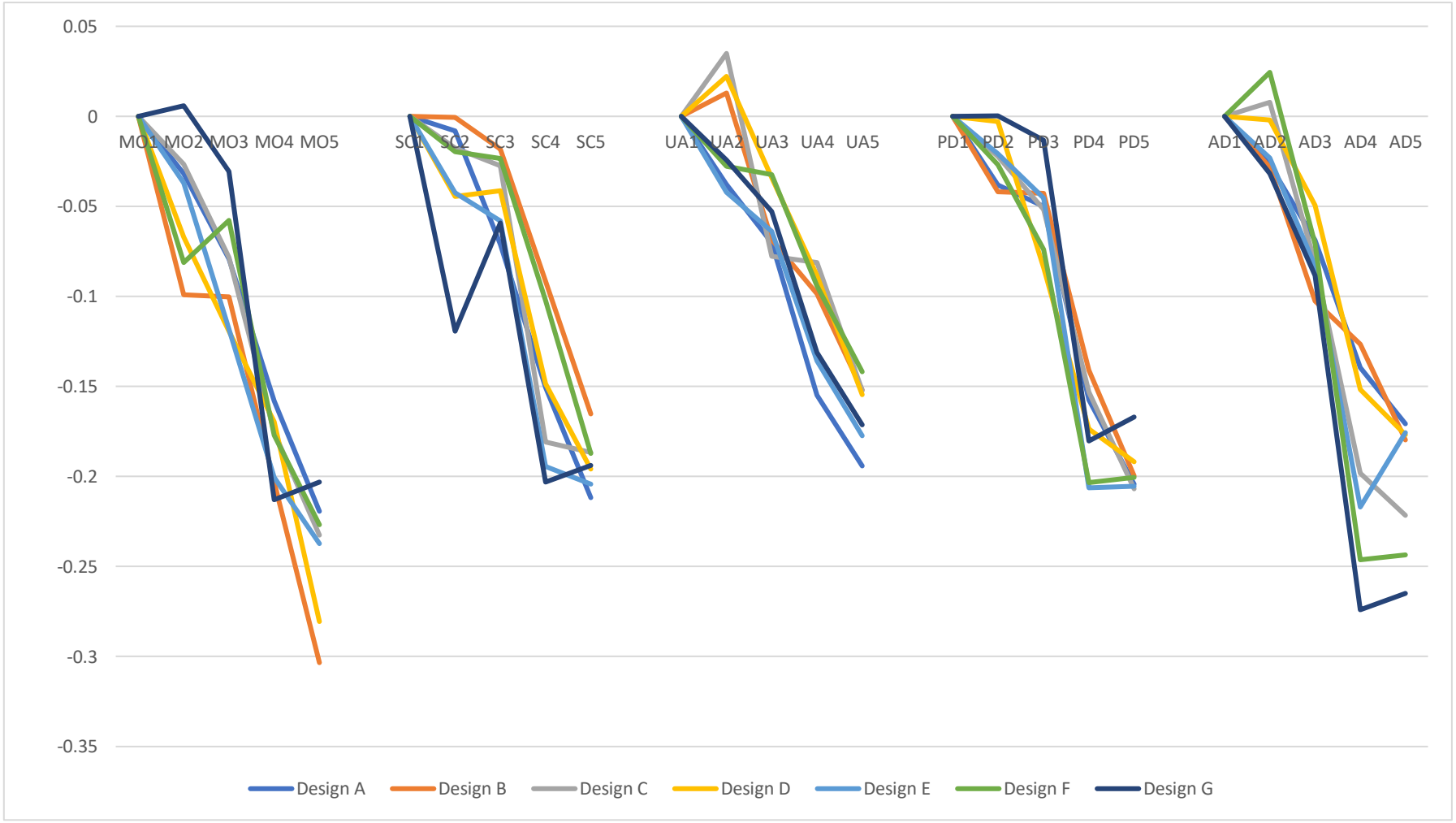


Figure 30: Parameter estimates scaled on the value of health state 55555 (non-overlap designs)

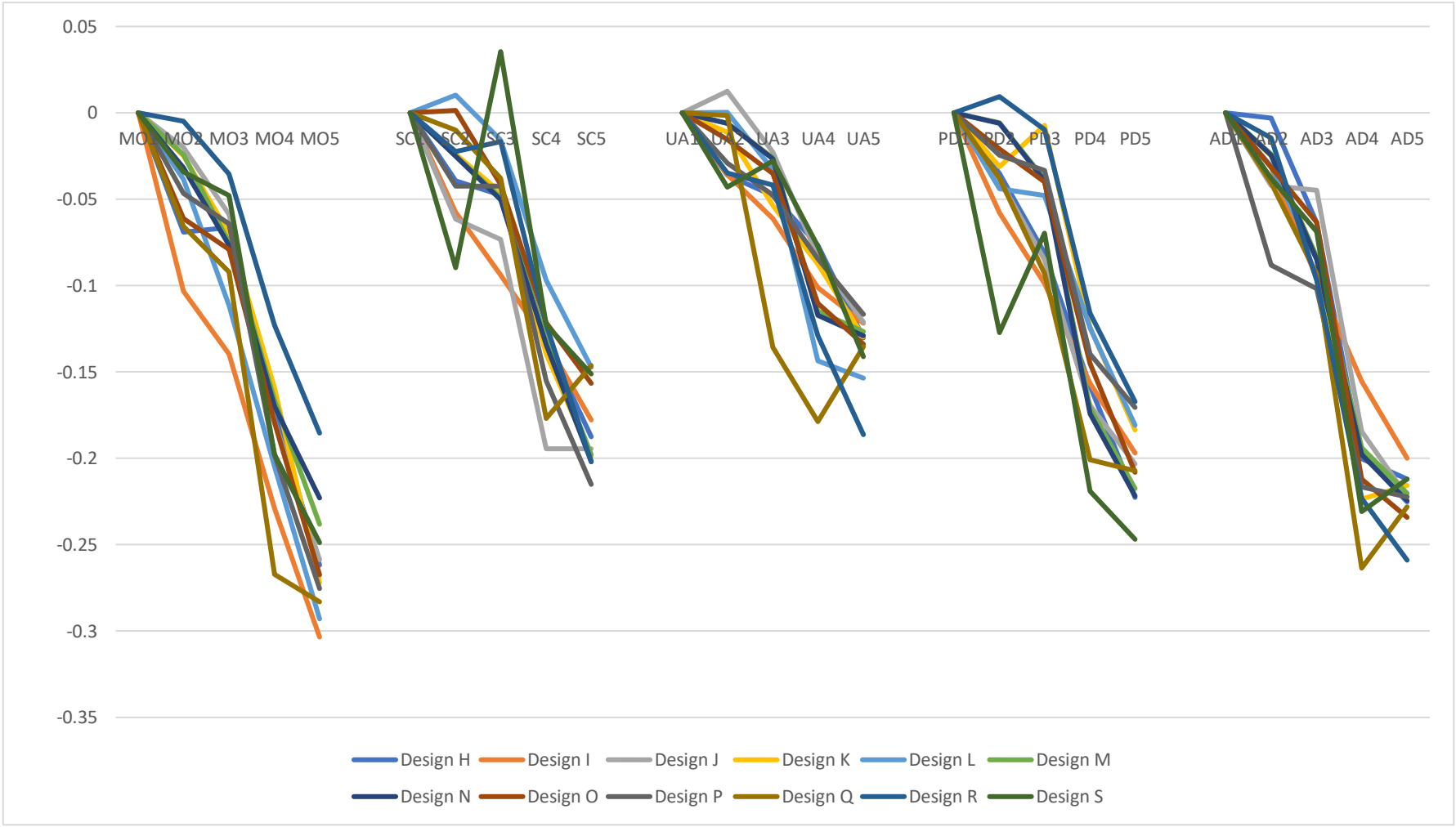


Figure 31: Overall magnitude of the scaled level 5 parameter for each dimension (overlap)

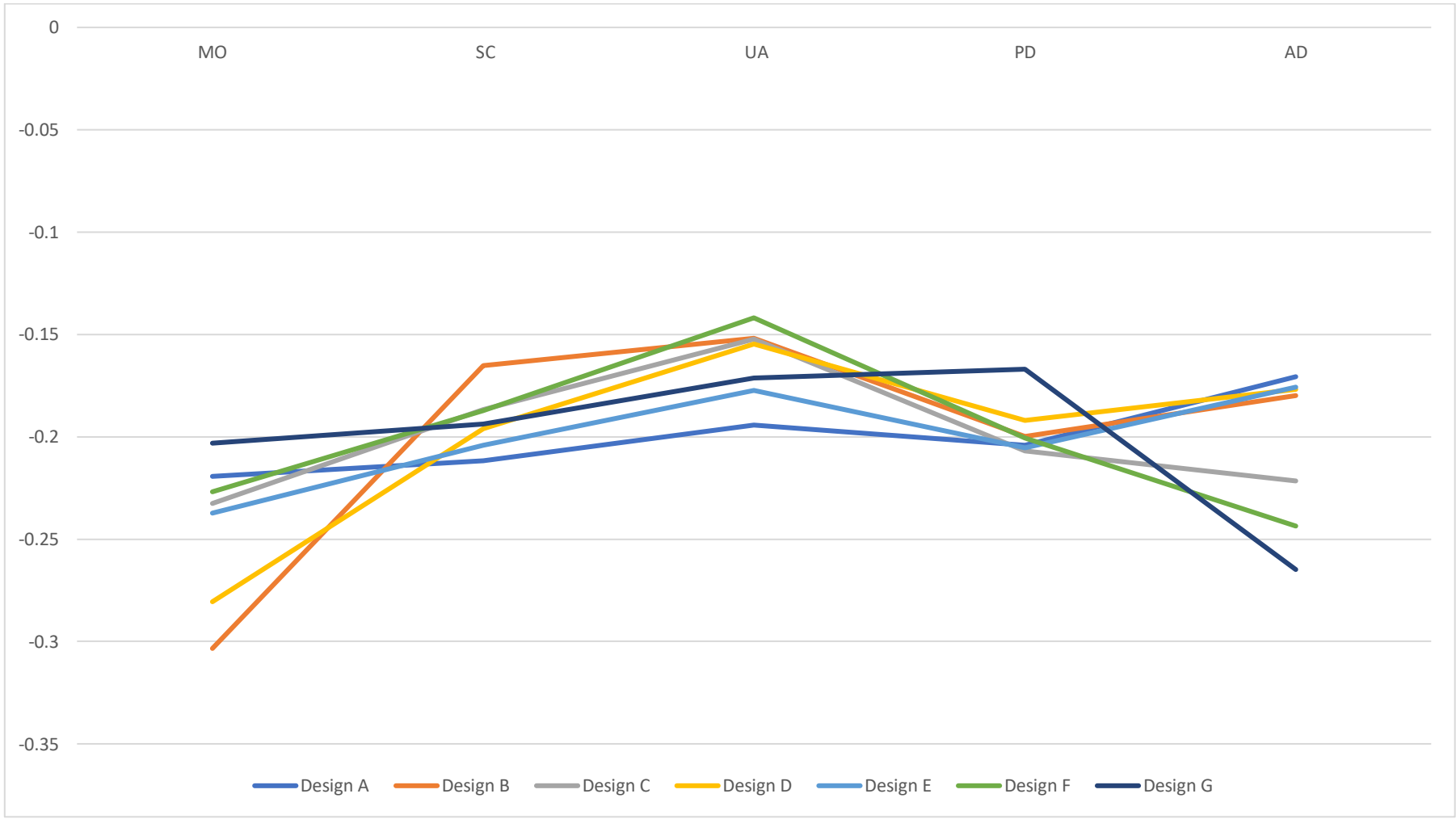
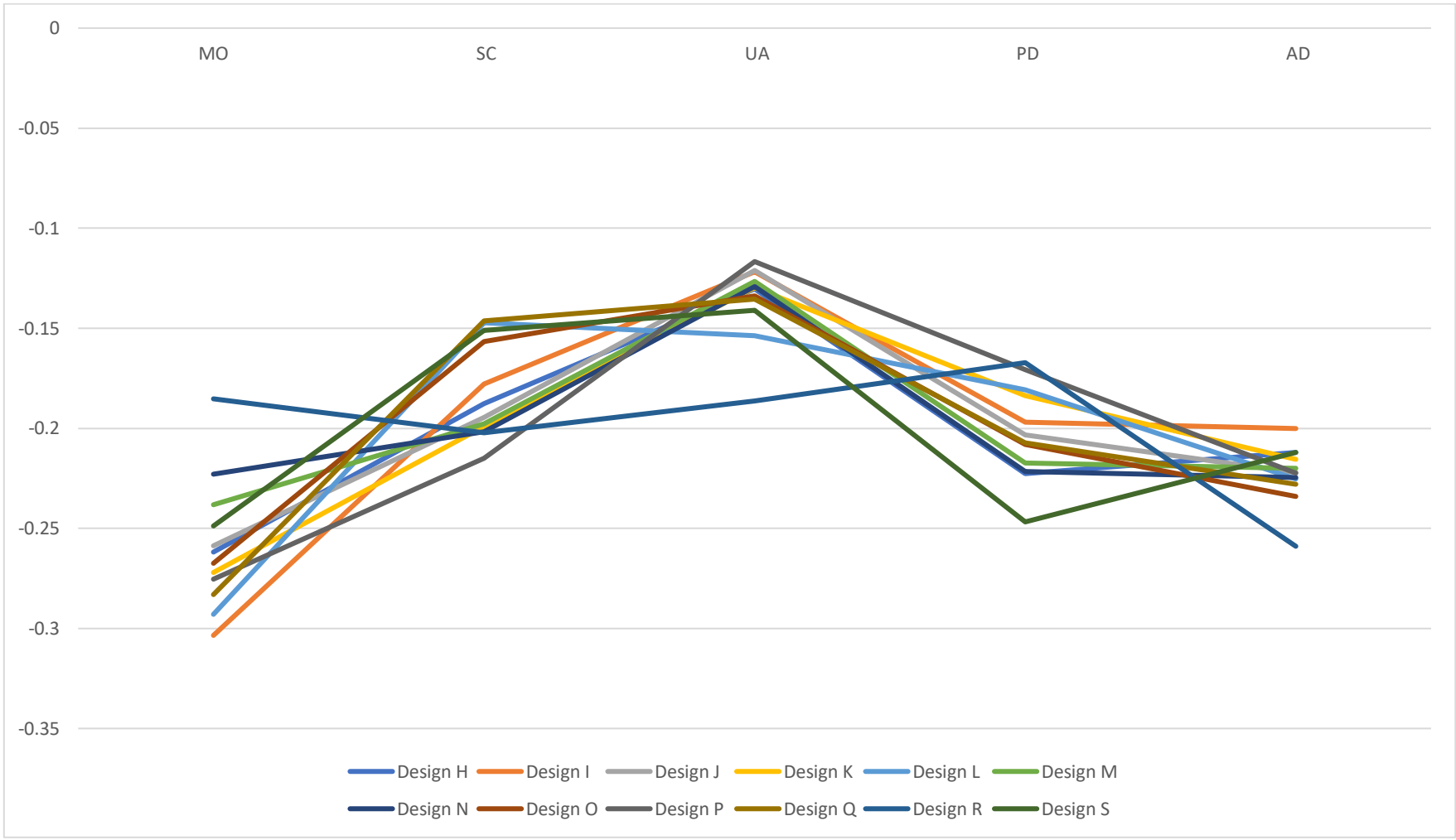


Figure 32: Overall magnitude of the scaled level 5 parameter for each dimension (non-overlap)



6.14.5. Comparing designs - Poolability

Table 60: Assessment of scale across the overlap and no overlap designs

	Model 49: Overlap pooled model		**Model 50: Non-overlap pooled model**	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE
MO2	-0.261***	0.043	-0.171***	0.022
MO3	-0.555***	0.048	-0.297***	0.040
MO4	-1.167***	0.068	-0.755***	0.041
MO5	-1.671***	0.084	-1.064***	0.053
SC2	-0.171***	0.041	-0.126***	0.021
SC3	-0.338***	0.046	-0.185***	0.022
SC4	-1.028***	0.063	-0.573***	0.034
SC5	-1.400***	0.077	-0.771***	0.042
UA2	-0.085*	0.041	-0.058**	0.020
UA3	-0.389***	0.043	-0.184***	0.022
UA4	-0.828***	0.055	-0.445***	0.029
UA5	-1.232***	0.069	-0.566***	0.033
PD2	-0.195***	0.043	-0.120***	0.020
PD3	-0.383***	0.048	-0.203***	0.023
PD4	-1.222***	0.070	-0.627***	0.035
PD5	-1.461***	0.079	-0.817***	0.043
AD2	-0.106*	0.045	-0.128***	0.020
AD3	-0.526***	0.049	-0.320***	0.024
AD4	-1.347***	0.074	-0.843***	0.043
AD5	-1.480***	0.078	-0.920***	0.047
Scale				
Design A	Baseline	n/a	n/a	n/a
Design B	-0.309***	0.063	n/a	n/a
Design C	-0.152*	0.064	n/a	n/a
Design D	-0.061	0.059	n/a	n/a
Design E	0.002	0.057	n/a	n/a
Design F	-0.250***	0.067	n/a	n/a
Design G	-0.095	0.062	n/a	n/a
Design H	n/a	n/a	Baseline	n/a
Design I	n/a	n/a	-0.263**	0.093
Design J	n/a	n/a	0.146*	0.062
Design K	n/a	n/a	0.140	0.100
Design L	n/a	n/a	0.093	0.063
Design M	n/a	n/a	0.170	0.099
Design N	n/a	n/a	0.031	0.069
Design O	n/a	n/a	0.252**	0.097
Design P	n/a	n/a	0.038	0.066
Design Q	n/a	n/a	-0.058	0.065
Design R	n/a	n/a	0.075	0.065
Design S	n/a	n/a	0.105	0.063

[a] Coefficient estimate; [b] standard error; p-values for the difference between the coefficient and baseline indicated by stars: *** 0.001, ** 0.01; *0.05; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; italics: inconsistent coefficients

The pooled model for the overlap designs is presented in **Table 60**. The scale parameter estimates for each design suggests that Designs B, C and F significantly differ from Design A. The LR statistic is 290.94, on 114 degrees of freedom (unrestricted parameters (7 x 20) = 140; restricted parameters (20 + 6) = 26). Therefore, the null hypothesis of scale homogeneity across designs is rejected. The pooled model for the non-overlap designs is also presented in **Table 60**. The scale parameter estimates for designs I, J and O significantly differ from Design H. The LR statistic is 406.37 on 209 degrees of freedom (unrestricted parameters (12 x 20) = 240; restricted parameters (20 + 11). Again, the null hypothesis of scale homogeneity across designs is rejected. This is indicative of design induced differences and means that the data from across the designs cannot be pooled.

6.14.6. Comparing designs - Assessing preference heterogeneity using latent class

Table 61 summarises the model performance indicators for latent class models with between two and five classes across the designs. In contrast to the conditional logit, scaling the latent class values on the worst health state is more difficult to interpret given the range of positive and negative values reported. Therefore, the latent values are used to compare class structures.

Table 61: Summary of latent class model performance

Design	Class size 2			Class size 3			Class size 4			Class size 5		
	LL	AIC	BIC	LL	AIC	BIC	LL	AIC	BIC	LL	AIC	BIC
A	-1,846	3,780	**3,921**	-1,804	3,745	3,963	-1,747	3,679	3,974	-1,719	**3,670**	4,042
B	-1,821	3,731	**3,871**	-1,760	3,657	3,873	-1,719	**3,622**	3,914	-1,697	3,627	3,995
C	-1,873	3,834	**3,972**	-1,818	3,773	3,988	-1,759	3,703	3,993	-1,733	**3,698**	4,064
D	-1,683	3,454	**3,594**	-1,636	3,409	3,627	-1,579	3,342	3,636	-1,541	**3,314**	3,685
E	-1,583	3,394	**3,254**	-1,522	3,180	3,397	-1,492	3,169	3,463	-1,458	**3,148**	3,518
F	-1,934	3,956	**4,096**	-1,883	3,903	4,119	-1,843	3,870	4,162	-1,806	**3,844**	4,212
G	-1,856	3,801	**3,940**	-1,808	3,752	3,967	-1,756	3,696	3,988	-1,723	**3,679**	4,047
H	-1,956	4,001	4,142	-1,890	3,917	4,134	-1,795	3,775	**4,069**	-1,756	**3,745**	4,116
I	-2,077	4,243	4,382	-1,990	4,117	4,333	-1,926	4,036	**4,327**	N/A	N/A	N/A
J	-1,771	3,631	3,771	-1,698	3,533	**3,749**	-1,643	3,471	3,763	-1,612	**3,457**	3,826
K	-2,092	4,273	4,411	-2,011	4,159	**4,373**	-1,967	4,119	4,408	-1,929	**4,091**	4,456
L	-1,819	3,726	3,866	-1,742	3,621	**3,837**	-1,686	3,557	3,849	-1,637	**3,506**	3,875
M	-2,146	4,380	4,518	-2,082	4,300	4,515	-2,003	4,190	4,480	-1,939	**4,110**	**4,476**
N	-1,887	3,863	4,002	-1,819	3,774	3,989	-1,750	3,685	**3,976**	-1,703	**3,639**	4,006
O	-2,176	4,441	4,580	-2,076	4,289	4,505	-2,034	4,252	**4,544**	-1,983	**4,198**	4,566
P	-1,883	3,854	**3,994**	-1,827	3,791	4,006	-1,762	3,708	3,999	-1,709	**3,651**	4,018
Q	-1,846	3,780	3,920	-1,739	3,614	**3,830**	-1,678	3,541	3,834	-1,616	**3,465**	3,834
R	-1,841	3,770	3,911	-1,779	3,695	3,912	-1,711	**3,606**	**3,900**	-1,687	3,607	3,979
S	-1,754	3,596	**3,736**	-1,696	3,528	3,744	-1,662	3,509	3,801	-1,604	**3,440**	3,808

The overlap designs consistently result in a two class latent class model being the preferred (as assessed by the BIC). The AIC results are unstable, suggesting that models with four and five classes are preferred. Given the BIC was consistent, the models with two classes were chosen for the seven overlap designs. The non-overlap designs are less consistent, with models with

two to five classes preferred according to the BIC, but with differences between the BIC recommendations. The BIC results were used to guide the choice of preferred model given its use in the earlier DCE work reported in this thesis. This meant that for the 12 non-overlap designs, one (Design M) was estimated with five classes, five (Designs H, I, N, O and R) were estimated with four classes, four (Designs J, K, L and Q) were estimated with three classes, and two (Designs P and S) were estimated with two classes.

The model performance indicators suggest that restricting the design in terms of imposing overlap may also limit the patterns of heterogeneity observed. The designs with no overlap are considerably more variable indicating that allowing all levels to vary may result in more diverse preferences within the choices made, where more information about heterogeneity can be gained from the profile comparisons within each choice set.

Figure 33 reports the two class models for the overlap designs, and Appendix 18 reports the demographic parameter estimates, with Class 2 as the baseline, and the class shares. Across all designs, there is one class (ranging between 46% and 80%) with generally strong and ordered preferences across the five dimensions. The other class demonstrates more evidence of disordering and coefficient estimates of a much lower magnitude. The respondents in this class are more likely to be older and generally report not having a long-term health condition. There is not strong evidence of a class pattern according to design feature. The disordered classes may not have strong preferences for the dimensions and levels included, or this could be evidence of lower task engagement (although this is difficult to interpret).

Figure 33: Latent class models with the lowest BIC (overlap designs)

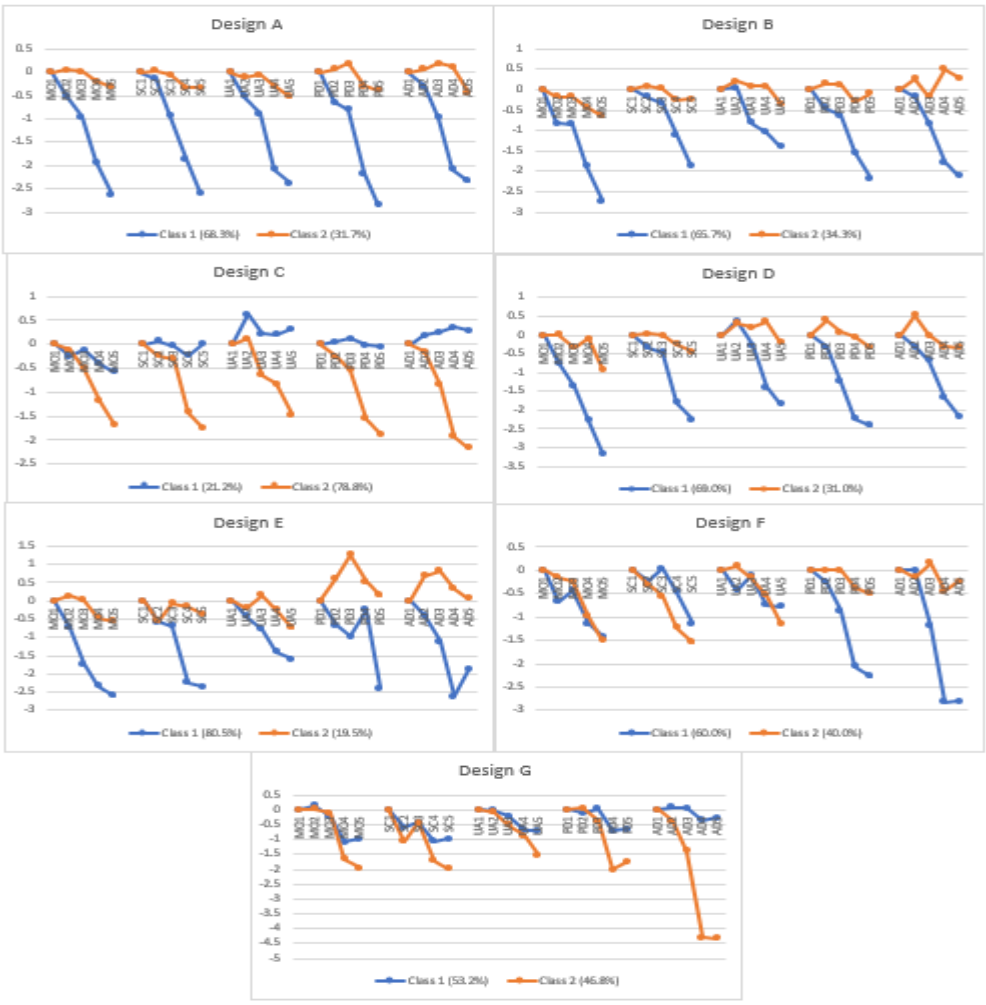


Figure 34: Latent class models with the lowest BIC (non-overlap designs)

Figure 34 reports the favoured class structures for the non-overlapping designs. Appendix 18 reports the demographic class memberships for each of the preferred models. As with the overlap designs, there is generally evidence for the identification of at least one class with strong and generally ordered preferences across the dimensions, with the designs with a two class structure (P and S) having similar characteristics to the overlap designs. The three class models for designs J, K, L and Q include classes that are generally ordered, but with different preference patterns across different dimensions. For example, there are classes with strong preferences to avoid a high level of problems with anxiety or depression, and pain or discomfort. Designs Q and M include a class with a disordered pattern of preferences with correspondingly large coefficients which are difficult to interpret (but may be due to instability in the latent class models).

6.14.7. Comparing designs - Assessing preference heterogeneity using mixed logit

Figure 35 to **Figure 37** display the results of the mixed logit analysis for the overlap and non-overlap designs, with the lines representing mean coefficient decrements, and the bars representing the standard deviation for each parameter, where the larger the standard deviation, the greater the heterogeneity for that parameter estimated for that design. There is evidence of more preference heterogeneity for the more severe levels of each dimension, and a mixed pattern for less severe dimensions. The results do not suggest that one type of design is more likely to produce models with clear evidence of preference heterogeneity than others (given that the samples completing each design are generally well matched in terms of background characteristics).

Differences across the designs were found when comparing the predicted probability of choosing option A with the observed frequency for each design in terms of percentage agreement. Regarding the overlap designs, Design A (generator developed) had the lowest level of agreement at 36% and Design E (SAS with Krabbe priors) had the highest at 58%. Regarding the non-overlap designs, Design I (also generator developed) had the lowest agreement at 22%, with Design M (Ngene swapping algorithm with Krabbe priors) having the highest at 70%.

Figure 35: Mixed logit models (overlap designs)

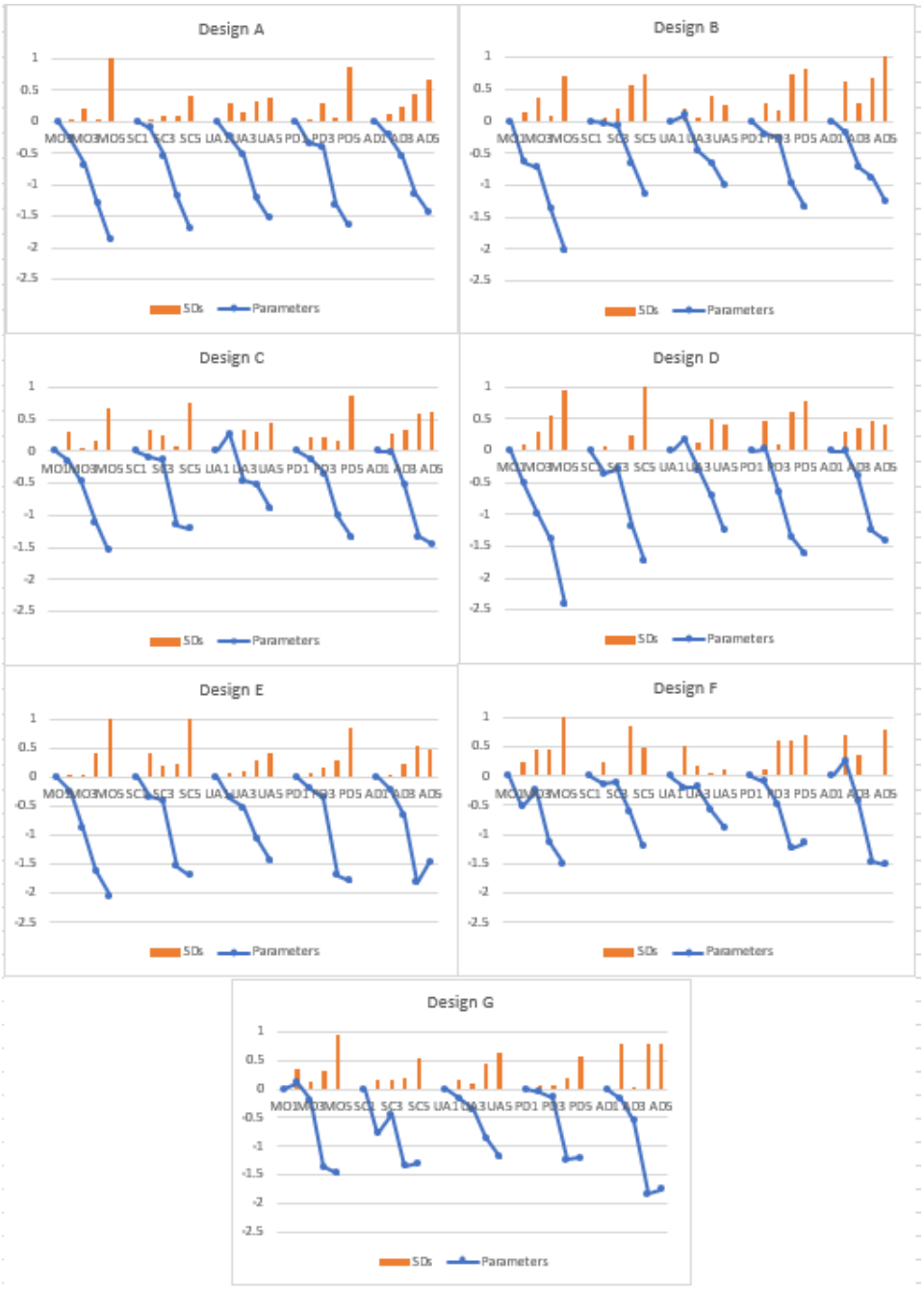


Figure 36: Mixed logit models (non-overlap designs H to M)

Figure 37: Mixed logit models (non-overlap designs N to S)

6.15. Summary and discussion

The results of DCEs to estimate health state values will be influenced by the design decisions made in the construction of the experiment. The aim of this study was to focus on a number of these features that are key to design construction. These were the design construction method, the level of overlap, the software that implements the selection mechanism, and the value of the priors used.

The results suggest that there is not one set of features that consistently produces better models in terms of the indicators studied in this paper. These indicators focused on key issues often assessed in DCE studies valuing PBMs, including the consistency of coefficient estimates and the precision of estimates. However, the poolability analysis suggests that there are still differences

in the scale of the designs despite matching the sample as far as possible in terms of demographics. There are advantages and disadvantages of the design features tested. Each feature will be considered in turn, and this is followed by study limitations and suggestions for further work, and conclusions with tentative recommendations for future DCE studies to develop value sets.

6.15.1. Overlap of severity levels

Enforcing overlap across dimensions within a choice set by fixing the severity level of a dimension to be the same across the two options within a choice set has grown in popularity to lessen the cognitive burden by lowering the amount of information to be considered [44]. However, enforcing overlap does mean that designs will have lower statistical efficiency for estimating the parameters. This could be an issue for studies with a small sample size. In this study, the seven designs with enforced overlap have bigger standard errors, but respondents reported that they were significantly less difficult than those without overlap based on a number of indicators. Taken together, these results demonstrate the trade-off between statistical and respondent efficiency. There are advantages to using statistically efficient designs in terms of the number of choice sets and respondents required. The disadvantage is that more difficult tasks may encourage respondents to adopt decision making short cuts. One such short cut is focusing on a subset of domains. However, differences in the use of decision making strategies was not apparent in responses to the follow-up question about considering the whole health profile whilst completing the tasks. However other completion strategies are possible (for example counting severity level indicators, or focusing on the most severe levels only). Increasing respondent efficiency may be more important for longer PBMs with more dimensions (such as the cancer specific QLU C10D [144]) or more complex descriptions (such as the SF-6D [79, 80] and the ASCOT [126]). In the online environment information about respondent behaviour or efficiency gained by overlap is difficult to measure beyond indicators such as the time taken, and answers to self-report questions.

The time taken based on the level of overlap is also potentially informative, but possibly contradictory. The fact that non-overlapping designs are completed faster on average may differ to what is expected, as the non-overlap design may be expected to take longer given the increased complexity. However it may suggest that respondents are simplifying the task for non-overlapping pairs by not attending to some dimensions. Where there are overlapping pairs it may simplify the task sufficiently that respondents can engage more easily. This interpretation

would imply that response time may not be a good measure of engagement. Qualitative work could be used to understand engagement based on overlap in more detail

6.15.2. Use of prior information

The use of priors in the design of DCEs has positive and negative impacts on the completion process and subsequent models produced, and design performance can vary based on the error in the priors [231]. Priors can either be non-informative or informative in the design process, and can be taken from the existing literature or previous studies (as in this comparison) or from pilot data. The benefits of using informative priors are that the design process can use pre-existing information about the size of the coefficients to maximise the information available from each choice set. The negative aspects are that using information to design the choice sets can make them too difficult for respondents and lead to more disordered coefficients, as was found in a previous design comparison paper valuing EQ-5D-5L [59]. However, there was not a clear difference between the consistency of the coefficients between the non-informative and informative priors used in this study. Furthermore, the results may be biased if inaccurate priors are used (which is possible given sample, design and study differences in earlier work) [257]. In this study the prior used in the design were not replicated in the field which could be due to these reasons. The argument for using non-informative priors is to present respondents with a wider range of comparisons with health states that are further apart in terms of severity to make the set of tasks easier. This may be particularly useful if larger samples are available to negate the efficiency gains, and theoretical design methods are favoured.

6.15.3. Theoretical or algorithmic approaches

Both theoretical and algorithmic based approaches have been used extensively in the development of DCE designs for the valuation of PBMs (see Section 2.9.4). The head-to-head comparison reported in this chapter has demonstrated that the theoretical generator developed construction approach employed here produced a generally lower number of inconsistent coefficients than the algorithmic based approaches, but this is only one indicator used. Both sets of designs seem to allow for identification of preference heterogeneity, particularly for the more severe levels of each dimension. Both sets of designs seem to be able to pick up preference heterogeneity at a similar level, particularly for the more severe levels. Algorithmic based approaches may be more appropriate if informative non-zero priors are planned or available for use in the study.

6.15.4. Impact of different software packages

Many software packages are available to generate designs, particularly for the algorithmic based approaches, and this study employed four that implement the same algorithmic approaches in different ways. The results suggest some differences in model performance across the software, where SAS and a user written implementation in R resulted in more inconsistencies. Ngene and Stata produce relatively well ordered models. Further work is required to understand and compare the different implementations of the same underlying algorithms across different packages.

6.15.5. Are certain designs better for certain models?

The range of models tested here, including those that impose preference homogeneity and allow for preference homogeneity provide the basis for an assessment of whether the characteristics of certain designs is more acceptable for certain model applications. The range of homogeneous and heterogeneous models tested here allow for an assessment of whether certain designs seem to be more valid with certain model applications. The conditional logit demonstrates differences in coefficient ordering and significance dependent on the design tested. The latent class models do seem to differ depending on the level of overlap, where designs with overlap on two dimensions always results in a two class model (that can be clearly interpreted) being preferred. The latent class models are less stable for the designs with no overlap, potentially demonstrating the extent to which a design strategy may influence the modelling of heterogeneity. The mixed logit models generally demonstrate a similar level of estimation precision, and more heterogeneity at more severe levels is apparent.

6.15.6. Study limitations and future research

This study has a number of limitations that could be important in the interpretation of the results, but also offer areas for future research. Although the coefficients were scaled using the estimated value of 55555, no method of anchoring onto the utility scale was included within or alongside the DCE. This limited the results as the values could not be directly compared on the QALY scale. It also meant that more complex designs such as those including continuous duration alongside categorical health state dimensions could not be compared. It should also be noted that the designs were developed for the MNL model. In future work, designs developed to estimate the mix logit model could be developed and compared.

Using the EQ-5D-5L as the descriptive system valued in this study has both advantages and limitations. The descriptive system is amenable to testing in this way given that the dimension descriptions are ordered (from no problems to extreme problems), there are a small number of dimensions, and the majority of respondents qualitatively understand the increasing severity levels. Therefore, it provides a useful set of descriptors to test coefficient ordering and heterogeneity. The applicability of the results to other measures could be questioned, but aspects of the results can be considered in regard to the valuation of wider and more complex QoL descriptive systems.

The informative priors used in this study were chosen as they were estimated from a DCE without duration of the EQ-5D-5L, and there was no clear pattern of differences across the designs with non-informative and informative priors. However, the appropriateness of the priors could be questioned given that they were sourced from the population of four countries, but applied in Australia. This may impact on the applicability of the comparison of the designs across the different priors, but it can be argued that it is relevant to compare any set of non-informative and informative priors to understand the impact of providing any prior information to inform the design. Alternative priors that were considered included values from an Australian DCE with duration study [57], but this was not used due to the specific methodological issues relating to the use of DCE with duration. Future work could pilot a design with zero prior to estimate a set of non-zero priors for use in the development of a further set of designs.

The overall sample size used for this study is at the high end of those collected in the application of DCE for health state valuation online (See section 3.3.4). However, as the sample was divided across 19 arms, the sample size is not large, and this could be a limitation in interpreting the results. This study could not report GMNL models given the small sample size, as exploratory modelling proved to be unstable. A larger sample could help understand whether certain design strategies are more amenable to complex models such as the GMNL. However, the sample size does provide indications of the direction of the findings.

There is also the potential for implausible combinations of EQ-5D-5L dimension levels. However, no dimension level combinations were excluded from the constructed designs developed in this study.

6.15.7. Conclusions and recommendations for study designs

In conclusion, the study reported in this chapter is the first to test a range of DCE design strategies used to value PBMs. The results provide information about a range of key decision factors in the development of designs, and will be informative for researchers developing valuation studies for newly developed measures of QoL.

The results of the study has implications for broadening the valuation of QoL, and allow for some initial recommendations for the future conduct of PBM valuation studies measuring both HRQoL and broader QoL outcomes using DCE. Given that respondents report that the overlap designs were easier to complete, future studies should impose a level of overlap, particularly when larger samples are available to negate the impact of imposing overlap on statistical efficiency. Overlap could also support ease of completion when complex QoL descriptive systems are valued, triplet profiles are used instead of pairs, or a duration anchor is included. Priors should be considered for use if there is an argument for them having a level of similarity with those expected from the sample. Dominant pairs and level balance do not seem to lead to substantial differences in the results and should be considered in light of the ease of task completion for respondents. There is not clear evidence to suggest that one type of theoretical or algorithmic design construction method is clearly superior to any other, although by their nature, random designs could lead to less ordered valuations depending on the designs selected. And if the inclusion of informative priors is preferred then an algorithmic design would be required. However overall the design construction method is a choice of the researcher.

7. Discussion

7.1. Summary

The overall research question of this thesis investigated how methods for the measurement and valuation of health and QoL can be used to inform the development of broader and more widely applicable instruments. The rationale for investigating this question was to understand how to increase the scope and applicability of utility values used in resource allocation decision making. There are two ways to increase the applicability of the utility values. First, it is possible to improve and refine the scope of what is measured, and second it is possible to improve the valuation process used to estimate utilities for use in decision making. Both of these research areas were investigated to answer the research question. To do this, three empirical studies and a structured review were conducted, with five aligned aims and objectives.

The results of the thesis suggest that it is possible to develop innovative broader measures of health and QoL, and for respondents to value the states that are described. This can be informed by adapting existing psychometric and preference elicitation methods to be fit for purpose for the assessment and valuation of broader constructs.

Regarding the measurement of outcomes, the results have demonstrated that broader concepts relevant to both health and wider QoL can be included in the same measurement framework using IRT approaches adapted from those used to assess narrower concepts, such as HRQoL. It was found that the dimensions included within existing PBMs frameworks can be extended to incorporate further broader dimensions of QoL. It was also found that PBM frameworks can be extended to provide further information for each dimension included in existing instruments whilst also providing preference-based information. This was defined as a 'layered' approach to measurement. The advantages, disadvantages and implications of this approach are described in detail below.

Regarding the valuation of outcomes to generate value sets, the appropriateness of using and adapting DCE methods for the valuation of broader measures of health and QoL, was established. The results also added to existing knowledge relating to the application of DCE methods for the valuation of health and QoL, in particular how to construct the experiment. Combining the evidence from the empirical work supports the development of a broader measure of health and QoL that are useful for decision makers in the allocation of scarce

healthcare resources. The results also support the further development of methods for both the measurement and valuation of health and QoL that can be used for this purpose.

In this chapter the key findings and issues raised by work conducted in this thesis are explained. First, the findings relating to broadening the measurement of QoL are highlighted, and then the implications of these are discussed. This is followed by a discussion of the issues raised for DCE as a valuation method specifically focusing on the valuation of broad areas of health and QoL, and the design of DCE studies to elicit accurate preferences. Leading on from this, the limitations of the work and further potential research areas are highlighted. Key questions raised by the research are then considered. Finally, the chapter concludes with reference to the overall research question posed at the beginning of the thesis.

7.2. Broadening the measurement of health and QoL

This thesis builds a case for the benefits of broadening how QoL is measured by proposing two possible approaches which are then investigated using IRT methods. The first approach proposed broadens the HRQoL measurement framework to include wider definitions of QoL measured by broader dimensions. This is a beneficial approach, as using a broader framework could improve measurement sensitivity in conditions and populations that impact wider areas of QoL than are included in HRQoL instruments.

The results of the empirical work demonstrated that the measurement of QoL should be broadened to improve the measurement of important outcomes in different conditions and populations. This is because there are important and diverse relationships between what is measured by different instruments developed to assess QoL. However not all of these are included in the resource allocation decision making process. The results established a dimension structure incorporating dimensions typically included in HRQoL measures, but also additional broader concepts of QoL such as mental health, physical functioning, pain, general activities, wider QoL (social care and capabilities), sleep, dignity, tiredness, role and social functioning, feelings, positive energy and mental health, and self-care incorporating related issues.

Four of these dimensions (mental health, physical functioning, pain, role and social functioning) focus on important HRQoL factors, on which impacts are commonly experienced as a result of many health conditions, and across many populations. Therefore these constructs are widely measured (taking different approaches) by both generic and condition specific measures. As an

example, they are aligned, and clearly map to, four of the five dimensions measured by the EQ-5D and five of the six measured by the SF-6D. These dimensions also cover the key domains included in many definitions of what 'health' is. For example the WHO definition [16] describes health as "a state of complete physical, mental and social wellbeing", and NICE [3] define HRQoL as "A combination of a person's physical, mental and social wellbeing; not merely the absence of disease". Therefore it could be argued that these four domains are essential for inclusion in any measure of health and QoL.

As a result of the inclusion of these dimensions in existing measures, improvements or worsening across them as a result of interventions and treatments will be reflected in the utility values used in the majority of decision making contexts. However, the nature of PBMs as concise measures of HRQoL with a limited number of dimensions could bias the allocation of resources towards interventions for conditions where change on these four common dimensions can be demonstrated at the expense of measuring the benefits of interventions with impacts on wider domains.

This leads to the question of what other domains of health and QoL could extend the scope and applicability of the utility values used in decision making. The additional seven dimensions identified in this thesis suggest areas for the extension of QoL measurement, at least within the framework of the measures included in this study. These extensions might include wider QoL (social care and capabilities), sleep, dignity, tiredness, feelings, positive energy and mental health, and self-care including related issues.

To examine these extensions further, it is worth considering the dimensions that do appear in either the EQ-5D or SF-6D. Self-care is a single dimension of the EQ-5D (assessing washing and dressing) that in this study groups with similar concepts from broader measures, for example appearance and feeding. In contrast, the SF-6D includes an item about bathing and dressing limitations that is incorporated into the PF dimension. This means that a direct values for self-care preferences cannot be elicited as it is included in the dimension covering physical functioning. This demonstrates that different approaches to the measurement of similar outcomes may change the characteristics of the values used in decision making. Issues regarding different approaches to how outcomes are measured are returned to below. Broader concepts may be considered when people respond to general self-care items (for example relating to looking after appearance), but these are not explicitly included in utility values. This assumption

could be assessed qualitatively by asking people what they are considering when answering the questions. Given that impacts on self-care are likely to be important for more severe conditions, and in the elderly, it can be argued that it is an important concept of QoL to measure if an instrument is developed to be applicable across several populations.

The additional SF-6D dimension measures energy. In the study reported in this thesis, energy is included as part of the positive energy and mental health dimension. That items assessing mental health and energy using positively worded items load together is important, as it suggests that the direction of the item wording is important. Energy is not directly measured by the EQ-5D, and research has suggested that additional dimensions assessing this could be important [112]. Linked to the positive energy item, a dimension measuring tiredness and fatigue using items worded in a negative direction was also found. Again this suggests that the way in which items and dimensions are worded is important. Many conditions impact on energy levels, and interventions can have both positive and negative effects on this. To broaden the measurement of QoL, assessing energy, tiredness and/or fatigue using items is important. Another broader dimension assesses sleep, which appears in neither the EQ-5D nor SF-6D, but is an important side effect of treatments and conditions, and often impacts other dimensions.

A dimension assessing a broad range of areas of wider QoL with no overlap with the HRQoL dimensions was identified. This demonstrates the need for the inclusion of broader domains that are not measured by those focused on HRQoL. The concepts included within the wider domain are diverse (for example ranging from feelings of control, love and security). In further development of QoL measurement these domains to be considered as a broader set of dimensions rather than one overall concept. Further psychometric work could understand the information provided by these domains in comparison to other QoL concepts to support the development of broader measures.

An additional dimension assessing positive feelings provides a different broader perspective to the measurement of mental health as it includes positively framed constructs such as happiness, friendship and support. This dimension did not fit with the negatively framed mental health domain that focused on common concerns such as anxiety and depression, which suggests that positively worded items frame mental health in a different way. This raises questions about the perspective that measures should take in the assessment of outcomes, and also how to incorporate these into the same instrument.

Finally a dimension measuring dignity was identified. This is an important concept impacting QoL, particularly, for example, in the elderly or in certain care situations. The measurement of dignity is challenging as it is a multifaceted concept. It may not be required for all populations and conditions, and should at least be considered in settings where dignity may be impacted.

The findings of these analyses raise a number of issues and questions linked to the measurement of health and QoL. First, the results demonstrate how approaches to measuring and describing similar and overlapping constructs of QoL differ. For example, instruments use single and multiple items to measure each dimension. This means that different levels of information about each domain are being elicited dependent on the perspective taken. Taking physical functioning as an example, the EQ-5D uses a single item to measure this, whereas the SF-36 uses ten items, of which three inform the SF-6D (including conceptualising self-care as a part of physical functioning).

Second, there are differences in the way that items are described. For example, different specific constructs are investigated (including walking, bending/kneeling, and activity level, for example). The response levels used (for example severity or frequency), and the direction of the item wording (for example asking about constructs in a negatively or positively framed way) also influences the characteristics of the information elicited, and the relevance of the items in different settings. For example consider the measurement of pain. Asking respondents about the frequency of experiencing any pain is different to asking them about the severity of the pain they experience.

Nevertheless, the results demonstrate how the domains measured in diverse ways interact with each other. For example, there is evidence that items using different approaches to measuring mobility and physical functioning are nevertheless measuring the same underlying construct, and therefore severity and frequency responses, for example, can be combined to elicit broader information about the impacts of a condition on a certain dimension. This was investigated further in the assessment of developing a layered approach to measurement that is discussed in more detail in the following section.

The results of the dimensionality assessment can be used to guide the future development of broader QoL instruments, and this is a key outcome of the thesis. This work adds to the literature

by informing what could be measured and how this could be done using existing measures as the basis for future developments. It provides a basis to inform broadening the measurement of QoL outcomes.

7.3. Developing a layered approach to measurement

The second approach tested whether the existing framework of PBMs focused on HRQoL can be used to provide further information for each dimension already included in the measure. As described above, this has been defined as a "layered" approach to measurement. Layer 1 is defined as the preference-based layer that elicits utilities, with Layer 2 providing more detailed information about each dimension. For example, Layer 1 would consist of the five EQ-5D dimensions, and Layer 2 would be a further set of items to provide more detail. The rationale for this approach was to investigate the development of an innovative method of measurement where both preference-based (Layer 1), and non preference-based profile information (Layer 2) could be elicited from people completing the instrument using innovative analysis approaches based on IRT. The work conducted in this thesis adapting IRT methods suggested that the approach is feasible, and a layered measure based on existing or newly developed items and domains could be developed.

The dimensions tested in this thesis demonstrate the feasibility of developing a layered approach to measuring QoL outcomes. The items operationalised as Layer 1 were found to be items sensitive to the central range of severity for the construct being measured. This means that they can validly be used as a general indicator of the domain, and directly link to items in Layer 2 assessing milder and more severe levels of the domain. Taking the physical functioning dimension as an example – the Layer 1 PBM item assesses severity of walking about on five levels which cover a broad underlying range of severity. The Layer 2 questions then assess more detailed constructs of physical functioning focused on different severities of impairment including, at the milder end, limitations with vigorous activities, and at the more severe end, limitations with climbing one flight of stairs. This demonstrates the range of information that can be obtained.

This approach has a number of potential benefits in the extending the measurement of health, and also for the usefulness of the data collected. PBMs by their nature are limited in the information that they can provide. This approach could broaden the applicability of the instruments across a variety of clinical decision making, research focused, and routine outcome

measurement settings. For example, a single instrument could provide a utility value for use in resource allocation decision making, and also detailed information for each dimension covering a wide range of constructs for use in clinical settings to inform treatments. Alongside this, the detailed information could be scored using an IRT based approach to allow for comparisons across populations. This would also allow for innovative approaches to administering the instrument such as CAT. Limits of the approach could be linked to the items available for inclusion, and potential differences in the usefulness of the approach across different dimensions (for example those including different numbers of items).

Of potential interest is the use of the EQ-5D-5L items as a preference-based layer in a preference and profile based approach. This approach could be justified for a number of practical reasons. The items are simple and broad ranging, and can be answered by the majority of people. It is also the most widely used PBM internationally, with value sets available in many countries that could still be applied to the Layer 1 data. The IRT results suggest that the EQ-5D dimensions provide information across the middle range of the latent severity scale, and therefore could act as general severity items for a calibrated set of more detailed questions relating to each EQ-5D dimension. This would enable the estimation of utilities and provide more data about the wider impacts of a condition on important dimensions of QoL.

The analysis also tested a novel application of the use of IRT methods to inform the validity and development of QoL outcome measures by applying the approach to a set of measures assessing a broad range of QoL constructs. As well as informing the potential development of a layered approach to measurement, the analyses conducted also allow for an investigation of adding information within an existing PBM framework. IRT methods have been demonstrated to have the ability to calibrate items from different instruments with different perspectives on measurement and methods for asking the questions on the same underlying severity scale. The item calibrations allow for an understanding of the information provided by the items within each dimension. This was done in the analyses reported in this thesis, and is in line with the approach employed in the development of PROMIS item banks [204]. However, it extends this to use IRT in a novel way to consider the development of a layered approach within the dimension structures tested.

7.4. Using IRT methods in the assessment of PBMs

The work described in this thesis applies IRT methods to test PBMs, and by doing this contributes to the literature by demonstrating the use of these methods in a setting where in the past it has been limited. One parameter IRT has been previously used in the development of condition specific PBMs from existing profile instruments, and in the assessment of established measures [85, 192, 193], and the two parameter model has been used to inform the development of a mental health specific PBM [194]. However, these IRT models have not been used to examine item pools measuring broad concepts of QoL with the aim to inform the further development of PBMs. The novel approach taken in this study extends the use of these methods for the assessment of outcome measures, and demonstrates that the approach is feasible for the examination of dimensionality and the testing of items within a broader item pool.

The work in this thesis builds on and confirms the findings of other work that has used various methodological approaches to factor analysis to understand the dimensionality of item pools taken from HRQoL measures. For example, Finch et al [258]) used principal-component analysis and CFA to establish the wider dimensionality of an item pool combining PBMs measuring HRQoL and measures of positive wellbeing. The authors found nine possible dimensions that they defined as psychological symptoms, physical functioning, pain, sleep and energy, satisfaction, cognition, relationships, hearing, and vision. The overlap in terms of the additional dimensions across the studies is apparent, even though the inclusion of some different instruments also explains some of the differences observed. Both studies included mental health, physical functioning and pain dimensions (which are identified as key domains of health included in measures of HRQoL). Broader dimensions identified in both studies included sleep and energy, which were combined as one dimension in Finch et al [258], but were identified as separate dimensions in this study. This could be due to the dimensionality assessment methods used, or the characteristics of the sample. Other dimensions identified across studies could be linked to the measures included.

The use of different dimensionality assessment methods resulting in similar outcomes adds a further level of robustness to the findings. Additional work applying IRT methods to understand the dimensionality of other item pools would help generate new knowledge regarding the relationship between dimensions of QoL, and assess how to broaden QoL measurement in a sample of different populations and patient groups completing overlapping outcome measures.

7.5. Findings related to the valuation of QoL

A structured review of the application of DCEs for health state valuation conducted as part of this thesis summarised information about how the methods have previously been used, and the methodological characteristics of the approaches taken. The empirical valuation studies built on this evidence to investigate how DCE methods can be applied to value broader measures of health and QoL. This provided evidence on how people value diverse QoL outcomes. This evidence can be translated for use in the development of utility value sets for broader measures for use in resource allocation decision making. The key methodological issue of how to construct experiments to administer DCEs for health state valuation was also investigated. This was done to inform the design process to use in future studies valuing PBMs measuring HRQoL, and other diverse domains of QoL. The issues raised by these two studies for the valuation of health and QoL are now discussed.

7.6. Using DCE to value broader QoL measures

The key finding of the empirical work valuing combined QoL outcomes on the same utility scale was that people are willing to trade-off across diverse concepts of QoL as they have different magnitudes of preferences for different QoL constructs. This has implications for decision making that relies on a narrow conception of a QALY that focuses on HRQoL, particularly when many health interventions affect both HRQoL and broader domains of QoL and wellbeing. This could mean that the benefits of interventions with impacts on the broader areas of QoL are incorrectly estimated. For example, an intervention with impacts primarily on relationships or control over daily life will not be accurately valued, and therefore change in QoL will not be observed, thus leading to bias in decision making towards interventions with impacts on HRQol. A relevant example from a PBAC decision relates to Icatibant for hereditary angioedema. It was noted by the PBAC that the benefits related to increased security and control from the availability of the treatment rather than the health gain from treatment of attacks. These are factors that were valued in this study alongside HRQoL, and were shown to be important in people's preferences. The inclusion of wider dimensions, such as security and control, in valuations could allow for these to be considered in decision making. To support this, the results of the empirical work found that control was valued as one of the most important QoL dimensions (as important as those measuring HRQoL).

The results also provided evidence that DCE is a feasible and suitable method to value descriptive systems including broader concepts of health and QoL. Internationally, there is

recognition that concepts beyond health are important in resource allocation, but this raises challenges in valuation. This study contributes to the literature tackling this challenge by developing methods that provide a feasible approach to developing value sets incorporating these broader concepts. The findings of this study relate specifically to the valuation of HRQoL and SCRQoL, but the methods used provide scope to expand such indices to produce values that are more sensitive to the impacts of care services across a wide range of patient groups and settings. For example, the preference relationship between HRQoL and wellbeing may be of interest, and the same framework developed in this study could apply to trade-offs between generic and condition specific dimensions of QoL to assess preferences in general and patient populations.

7.7. Testing design construction approaches

The results of the final empirical study provide data regarding the impact of DCE design construction methods on the values elicited for health states. The study built on the existing literature reported in the structured review, and included design approaches that have been previously used for health state valuation, and also approaches that have not previously been applied. To the author's knowledge, it is the largest comparison of designs for the specific valuation of health states conducted in the literature to date.

The results suggested that there was not a clearly superior approach to the construction of a DCE for the purpose of health state valuation. However, the results do provide crucial information about a number of methodological choices that are required in the development of DCE designs. Therefore they provide data to support the development of further valuation protocols applying DCE to value health and broader QoL outcomes.

The first methodological choice the results inform regards the use of attribute level overlap (where severity levels are held to be the same for a certain number of dimensions). The results found that the overlap designs included in the study were easier for respondents to complete. Therefore they suggest that future studies should consider imposing overlap of attribute levels to aid ease of completion when complex QoL descriptive systems are valued. However the number of dimensions on which to impose overlap requires further testing.

The second area of design that can be guided by the findings is the use of priors to inform the design. The results suggested that both informative and non-informative priors can produce

logically ordered and comparable values, and there was not clear evidence to suggest that one type of theoretical or algorithmic design construction method systematically performed better. Therefore, priors should be considered for use if there is an argument for the values used having a level of similarity with those expected from the sample. For example, using priors from an Australian sample when constructing a design that will be applied to value an instrument in Australia.

This study established that there is no clearly superior approach, and the design used can be informed by the evidence for the different design features tested in this study, and how they can influence the characteristics of the values produced. The results therefore provide guidance that will allow the designer of the experiment to choose between different design features for the development of DCE studies for the valuation of health and QoL.

7.8. Broadening the measurement and valuation of QoL using existing measures

A key feature of the empirical work conducted in this thesis is the use of existing measures to investigate broadening the measurement and valuation of health and QoL. This provides one approach to investigating the research question, and has a number of advantages. For example, it provides a basis to test methods and the further development of measures using instruments that have already undergone extensive development and validation. To test the elicitation of values, existing measures are known provide dimensions that are amenable to valuation. However, the use of existing QoL instruments also has disadvantages as it limits the wider applicability of the findings to other measures and datasets.

Regarding testing the broader measurement of QoL, the dimension structure identified in this study is limited to the instruments included, and other dimensions might have been identified if different measures had been included. This means that the applicability of the results to newly developing measures such as the E-QALY [259] requires further investigation. However, the measures were chosen to cover a range of QoL concepts and did provide clear information about how additional instruments extend the standard HRQoL frameworks. The analysis could be repeated including either newly developed or other existing measures, using datasets where multiple outcomes have been collected [111].

7.9. Implications of the findings

The results of this thesis have implications for the use of existing PBMs in resource allocation decision making. They also have implications for the future development and valuation of QoL measures and inform a number of questions of importance for the research area and wider policy issues. These questions are posed and possible answers discussed below.

7.9.1. What do the results mean for the concept of QoL, and decision making based on QoL?

In this thesis, QoL has been described as a multifaceted concept that encompasses, and is impacted by, broad interrelated domains of an individual's life. The results of the empirical work in this thesis focused on measuring QoL outcomes support this descriptive framework. They show that the concept of QoL includes domains of physical health, emotional health, social functioning and relationships, and social outcomes that add to an overall interrelated description of broad QoL.

The valuation work focused on the interaction, and preference relationship between, two subdomains of this overall conceptualisation (HRQoL and SCRQoL). The empirical valuation work supports the assertions made about the interaction of different types of QoL, as the results demonstrate that people express different preferences for, and therefore trade between, these outcomes. This suggests that different areas of QoL, and population preferences for these domains, need not be considered separately, and can be conceptualised on the same utility scale. Therefore a unified approach to linking QoL outcomes on the same scale is needed, and this thesis provides a template for doing this. This would support the fairer allocation of health care resources across diverse medical conditions, interventions and patient populations that is fundamental in achieving better health outcomes for the population. Fair allocation of resources would also need to consider how priorities might change if a new framework was advocated. For example, interventions with a health care impact only might be deprioritised (depending on the magnitude of preferences for health domains in comparison to broader domains). Further work needs to investigate the implications of changing the values used in practical terms.

It could also be hypothesised that other domains of QoL not included in the empirical work, such as financial wellbeing and living situation, are also important drivers of QoL. Conceptualising QoL as a single unifying concept with interacting subdomains means that these concepts could also be used to inform public policy in many different areas. For example, the measurement and

valuation of QoL could be used to inform policy in housing, transport and the environment as well as health and social care. This would require sector specific measures to be developed, and an understanding of how these broader domains interact with health and other QoL outcomes. To achieve fair allocation of resources, we need to measure all of the QoL outcomes that matter to the population, and understand population preferences across these different areas of QoL. This thesis provides methods to broaden the scope of QoL outcomes used in diverse decision making contexts.

7.9.2. Can decision makers use existing PBMs with confidence?

This research has added to the literature demonstrating that widely used existing PBMs such as the EQ-5D and SF-6D do not include all of the dimensions of importance to patients [112]. This is partly driven by the typically preferred structure of PBMs where a short measure amenable to valuation for the development of value sets are required. There is evidence to suggest that the EQ-5D and SF-6D are psychometrically valid in many health areas, and do provide utilities that can inform decision making for health care. Decision makers need to be aware of the health areas where the measures have a level of validity, and where supporting evidence is not available.

The results of this thesis demonstrated that there are areas of health care and certain populations (for example the elderly) where health and social care are closely related, where these measures do not have the same level of validity. In these settings the standard HRQoL value sets may not reflect all of the concepts of importance, as key domains are not explicitly measured, or included in the values used in decision making. This is where a broader instrument may be useful, and there are real world examples from PBAC decisions that are worth examining in light of this. For example, PBAC noted the importance of social and psychological impacts for Poly-L-Lactic Acid for facial lipoatrophy were not captured by the SF-6D (indicating potential measurement insensitivity). It might be possible that a broader measure of QoL would demonstrate an increased level of sensitivity.

Across international decision making jurisdictions, different guidelines regarding the sources of utilities and values are used, and this would impact the use of broader measures. For example, NICE prefers values to be based on the UK EQ-5D-3L value set as the primary outcome [29], which could restrict the use of broader measures. In contrast, the Australian PBAC policy is to accept values from a wider range of sources as long as the source is justified. Therefore, broader

measures could potentially be justified as a source of values in decision making in Australia. Broader measures could also be informative outside of reimbursement decision making, for example in the routine collection of outcomes data and in clinical settings.

7.9.3. Should a broader approach to measuring QoL be advocated?

The results of this thesis suggest that for a more inclusive and holistic approach to measurement, a broader approach should be advocated. A strong argument can be made for an instrument including a broad range of QoL domains. This would lead to a more informed resource allocation process across different types of interventions and broader population groups.

7.9.4. If a broader measure of QoL was developed, what format should it take?

There are multiple forms new approaches could take and these are informed by the work conducted here, and also other research internationally [200,204]. The results of this thesis suggest that using up-to-date IRT methods for measure development is an informative way to understand the measurement characteristics of the dimensions and items included. There is also some support for adapting existing instruments by broadening what is measured. Incorporating a layered approach to measurement is advocated to increase the usefulness and wider applicability of any future measure.

7.9.5. Are DCE's an acceptable method for the valuation of health and QoL?

Past work using DCEs to value QoL (reported in Chapter 3) has generated evidence supporting the acceptability of the methods for estimating value sets. The work conducted in this thesis also builds on this to demonstrate that DCE can be used to value broader dimensions of QoL, and multiple methods of constructing a DCE are valid for the estimation of value sets. They are also practically valid as they can be administered online to large representative samples of the population, as demonstrated in the two studies reported here. However, the results suggest that the values produced are susceptible to different methodological choices that need to be understood for a given study or approach. Also, it is difficult to determine DCE data quality which could limit its wider acceptability. In the design comparison work, a repeat choice set was used to assess consistency, and there was some evidence of respondent not providing consistent answers. Further work should develop methods for assessing DCE data quality to improve the wider acceptability of the approach.

DCE values also differ to those estimated using other valuation methods (such as the TTO). For example, DCE values often produce a wider range of utility values, and more states valued as worse than dead (but again this is specific to the valuation method used). However, as there is no gold standard valuation approach, and we do not know what respondents' actual values are, but can only elicit these using indirect methods, the comparison is flawed.

7.9.6. What questions should decision makers and researchers ask when assessing the results from a PBM?

It is important for researchers and decision makers to understand the characteristics, advantages and limitations of both the descriptive systems and value sets of PBMs. For researchers, selecting the right measure will facilitate meaningful data collection and analysis, and for decision makers it is important to know how the concepts measured and the values applied can influence their decisions. Users should be aware which concepts are measured by different instruments, as this gives an indication of the types of impacts the measure and the value sets will be sensitive too. This thesis provided information to assess the relationship between broad measures developed for different purposes to demonstrate both overlap and divergence.

Regarding valuation, it is important for users to understand how the valuation approach used influences the value set characteristics. For example, different implementations of a DCE can lead to a wider or shorter value range (as demonstrated in the empirical comparison of designs), and DCE values can differ to those elicited using iterative approaches such as TTO. This also means that decision makers need to understand how the characteristics of value sets impact the results on which they make decisions.

7.10. Limitations of thesis and suggestions for further research

The research conducted in this thesis has a number of limitations that lead to associated opportunities for further research. As described in Section 7.6 the use of existing QoL instruments to assess the measurement and valuation of broader QoL has benefits but also limitations. This criticism could be answered by repeating the analysis on items from newly developed measures, or other existing generic and condition specific instruments in datasets where multiple outcomes were collected.

For the measurement assessment study, the population was recruited to represent some common conditions, and also include the general population, and therefore the results are limited to the groups included in this study. The use of an online panel could be criticised given that the conditions are self-reported rather than diagnosed. It would be beneficial to collect QoL information from patient groups, but this may have difficulties in terms of recruitment, and the burden of completing many questions for individuals from certain populations.

It is also worth noting that although IRT and DCE methods have been demonstrated to provide information about how to broaden the QALY, they are only one set of approaches to understanding how to do this. Therefore, the results need to be considered alongside other methods such as qualitative work with patients directly investigating the areas of health and QoL that are important and relevant to them. There are also other methods for the identification and assessment of dimensionality (such as CFA) that could be used alongside the approaches reported in this thesis to extend the evidence regarding the dimension structure.

A general area of potential limitations pertaining to the three empirical studies conducted in this thesis, but in particular to the DCE studies, is the robustness of online data collection, and the impact this has on data quality. Although there are indicators of 'valid' data that are generally accepted (e.g. for example sufficient time taken, not always selecting the same option, and responses to logic checks), DCE researchers have not yet fully established what a 'valid' or 'invalid' DCE response is. To improve online data collection, researchers should examine what a valid response is, and the extent to which DCE tasks are 'cognitively challenging'. This is an area of further research that requires the consideration of each stage of a DCE (from constructing the experiment and designing the choice sets through to testing mode of administration issues, and implementation issues such as the source of respondent recruitment, and respondent behaviour during the survey) to understand sources of bias on responses.

Regarding robustness and data quality, a particular area of concern is that it is difficult to control the environment in which the survey is completed, or collect information about this from respondents. We also do not know how different environments, or choices made in the survey design process, impact the responses given. Thus there is a need for further work to understand robustness and data quality in a number of ways.

First, it is possible to develop and test questions that ask respondents to report on the environment in which the survey was completed. However this could be influenced by self-report accuracy issues. Second, to understand environmental impacts on responses, it would be possible to conduct controlled experiments varying the environmental stimuli. Third, to understand survey design issues on response, further work investigating choice set presentation, or designing surveys to increase attention levels (for example imposing a minimum time for each task that is built into the survey rather than using observed time for each task for post hoc exclusions), or limiting screen size for which completion is allowed. Fourth, it might be possible to test controlling the environment in a number of ways, for example, requesting that respondents complete the survey at home only, or at a certain time of day. Finally, logic checks such as repeat and dominated tasks have a place in understanding completion and data quality, and should be considered in the design process.

The impacts of extending the QALY framework on previous resource allocation decisions have also not been tested using available data, so the practical implications for decision making have not been identified. This would be a natural extension to understand the implications of broadening measurement frameworks in more detail. Further work could consider the impact of using an instrument with a value set based on combined outcomes on previous decisions using existing clinical data.

7.11. Conclusions

The overall research question of this thesis asked how methods for the measurement and valuation of health and QoL can be used to inform the development of broader and more widely applicable instruments. The results of the thesis suggest that it is possible to adapt existing psychometric and valuation methods to develop broader measures of health and QoL, and value them to elicit sensitive and accurate utility values.

Regarding measurement an innovative approach to using IRT that combined multiple measures and outcomes was used, and the results demonstrate that this can inform potential ways to broaden the measurement of QoL. Regarding valuation, the results suggest that DCE can be used to value diverse QoL outcomes on the same scale, and different design strategies are acceptable for developing value sets. This work is innovative as it is the first attempt to value diverse outcomes on the same utility scale within the same choice set framework. It also provides the largest comparison of DCE designs for health state valuation purposes.

The results of this thesis can be used to inform the development and valuation of broader measurement systems. They can enhance the applicability, and increase the scope, of QALY values used in decision making, and therefore have implications for the allocation of scarce health resources.

8. Appendices

8.1. Appendix 1: HUI classification systems

Tables 62 and **63** display the HUI-2 and HUI-3 descriptive systems respectively

Table 62: The HUI-2 descriptive system

Dimension	Level	Description
Sensation	1	Able to see, hear, and speak normally for age
	2	Requires equipment to see or hear or speak
	3	Sees, hears, or speaks with limitations even with equipment
	4	Blind, deaf, or mute
Mobility	1	Able to walk, bend, lift, jump, and run normally for age
	2	Walks, bends, lifts, jumps, or runs with some limitations but does not require help
	3	Requires mechanical equipment (such as canes, crutches, braces, or wheelchair) to walk or get around independently
	4	Requires the help of another person to walk or get around and requires mechanical equipment as well
	5	Unable to control or use arms and legs
Emotion	1	Generally happy and free from worry
	2	Occasionally fretful, angry, irritable, anxious, depressed, or suffering night terrors
	3	Often fretful, angry, irritable, anxious, depressed, or suffering night terrors
	4	Almost always fretful, angry, irritable, anxious, depressed
	5	Extremely fretful, angry, irritable, anxious, or depressed usually requiring hospitalisation or psychiatric institutional care
Cognition	1	Learns and remembers school work normally for age
	2	Learns and remembers school work more slowly than classmates as judged by parents and/or teachers
	3	Learns and remembers very slowly and usually requires special educational assistance
	4	Unable to learn and remember
Self-Care	1	Eats, bathes, dresses, and uses the toilet normally for age
	2	Eats, bathes, dresses, or uses the toilet independently with difficulty
	3	Requires mechanical equipment to eat, bathe, dress, or use the toilet independently
	4	Requires the help of another person to eat, bathe, dress, or use the toilet
Pain	1	Free of pain and discomfort
	2	Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities
	3	Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities
	4	Frequent pain; frequent disruption of normal activities. Discomfort requires prescription narcotics for relief
	5	Severe pain. Pain not relieved by drugs and constantly disrupts normal activities
Fertility	1	Able to have children with a fertile spouse
	2	Difficulty in having children with a fertile spouse
	3	Unable to have children with a fertile spouse

Table 63: The HUI-3 descriptive system

Dimension	Level	Description
Vision	1	Able to see well enough to read ordinary newsprint and recognise a friend on the other side of the street, without glasses or contact lenses
	2	Able to see well enough to read ordinary newsprint and recognise a friend on the other side of the street, but with glasses
	3	Able to read ordinary newsprint with or without glasses but unable to recognise a friend on the other side of the street, even with glasses
	4	Able to recognise a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses
	5	Unable to read ordinary newsprint and unable to recognise a friend on the other side of the street, even with glasses
	6	Unable to see at all
Hearing	1	Able to hear what is said in a group conversation with at least three other people, without a hearing aid
	2	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people
	3	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people, with a hearing aid
	4	Able to hear what is said in a conversation with one other person in a quiet room, without a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	5	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	6	Unable to hear at all.
Speech	1	Able to be understood completely when speaking with strangers or friends
	2	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well
	3	Able to be understood partially when speaking with strangers or people who know me well.
	4	Unable to be understood when speaking with strangers but able to be understood partially by people who know me well
	5	Unable to be understood when speaking to other people (or unable to speak at all)
Ambulation	1	Able to walk around the neighbourhood without difficulty, and without walking equipment
	2	Able to walk around the neighbourhood with difficulty; but does not require walking equipment or the help of another person
	3	Able to walk around the neighbourhood with walking equipment, but without the help of another person

	4	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood
	5	Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood
	6	Cannot walk at all
Dexterity	1	Full use of two hands and ten fingers
	2	Limitations in the use of hands or fingers, but does not require special tools or help of another person
	3	Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person)
	4	Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools)
	5	Limitations in use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools)
	6	Limitations in use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools)
Emotion	1	Happy and interested in life
	2	Somewhat happy
	3	Somewhat unhappy
	4	Very unhappy
	5	So unhappy that life is not worthwhile
Cognition	1	Able to remember most things, think clearly and solve day to day problems
	2	Able to remember most things, but have a little difficulty when trying to think and solve day to day problems
	3	Somewhat forgetful, but able to think clearly and solve day to day problems
	4	Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems
	5	Very forgetful, and have great difficulty when trying to think or solve day to day problems
	6	Unable to remember anything at all, and unable to think or solve day to day problems
Pain	1	Free of pain and discomfort
	2	Mild to moderate pain that prevents no activities
	3	Moderate pain that prevents a few activities
	4	Moderate to severe pain that prevents some activities
	5	Severe pain that prevents most activities

8.2. Appendix 2: AQoL-8D classification system

Table 64 displays the AQoL-8D classification system

Table 64: How the AQoL-8D items contribute to the descriptive classification

Dimension	Item
Independent living	How much help do you need with jobs around your place of residence (e.g. preparing food, cleaning, gardening)? How easy or difficult is it for you to get around by yourself outside your place of residence (e.g. to go shopping, visiting)? How easy or difficult is it for you to move around (using any aids or equipment you need e.g. a wheelchair, frame or stick)? How difficult is it for you to wash, toilet, dress yourself, eat or care for your appearance?
Pain	How often do you experience serious pain? How much pain or discomfort do you experience? How often does pain interfere with your usual activities?
Senses	How well can you see (using your glasses or contact lenses if they are needed)? How well can you hear (using your hearing aid if needed)? How well do you communicate with others (e.g. talking, signing, texting, being understood by others and understanding them)?
Happiness	How content are you with your life? How enthusiastic do you feel? How often do you feel happy? How often do you feel pleasure?
Mental health	How often do you feel depressed? How often do you have trouble sleeping? How often do you feel angry? Do you ever feel like hurting yourself? How often did you feel in despair over the last seven days? How often did you feel worried in the last seven days? How often do you feel sad? Do you normally feel calm and tranquil or agitated?
Coping	How much energy do you have to do the things you want to do? How often do you feel in control of your life? How much do you feel you can cope with life's problems?
Relationships	How much do you enjoy your close relationships (family and friends)? How satisfying are your close relationships (family and friends)? How often do you feel socially isolated? How often do you feel socially excluded or left out? How happy are you with your close and intimate relationships? Does your health affect your relationship with your family? Does your health affect your role in your community (e.g. residential, sporting, church or cultural activities)?
Self-worth	How much of a burden do you feel you are to other people? How often do you feel worthless? How much confidence do you have in yourself?

8.3. Appendix 3: Structured review search terms

The search terms used to identify published papers reporting studies using DCE methods to value health states were as follows:

- Preference-based measure AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- EQ-5D AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- Euroqol and (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- SF-6D AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- (Multi-attribute utility instrument OR MAUI) AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- Utility measure AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- Health-related quality of life AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- Quality of life AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- Preferences AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- Health state valuation AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)
- Valuation AND (Discrete Choice Experiment(s) OR DCE OR conjoint analysis)

8.4. Appendix 4: CREATE checklist for reporting valuation studies

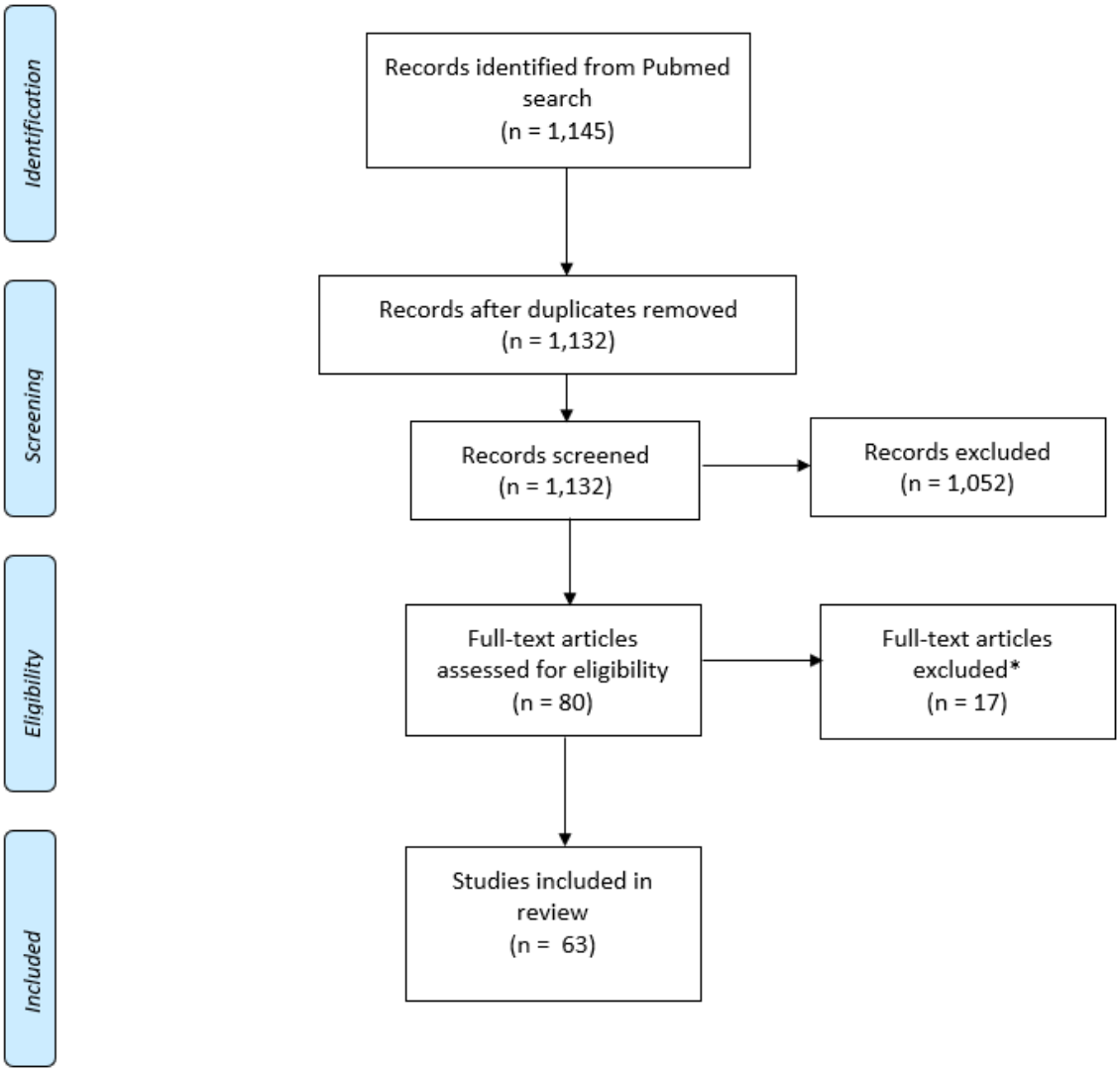
Table 65 displays the 'Checklist for Reporting Valuation Studies of Multi-Attribute Utility-Based Instruments' used to assess the quality of the DCE studies included in the structured review.

Table 65: CREATE checklist

Descriptive system			
1	The attributes of the instrument are described	□	□
2	The number of levels in each attribute of the instrument is described	□	□
Health states valued			
3	The approach to selecting health states to be valued directly is explained	□	□
4	The number of health states valued per respondent is stated	□	□
5	Method(s) of assigning the health states to respondents are stated	□	□
Sampling			
6	Sample size/power calculations are stated and rationalised	□	□
7	Target population is described	□	□
8	Sampling method is stated and rationalised	□	□
9	Recruitment strategies are described	□	□
10	Response rate is reported	□	□
Preference data collection			
11	Mode of data collection is stated	□	□
12	Preference elicitation technique(s) are described	□	□
Study sample			
13	Reasons for excluding any respondents or observations are provided	□	□
14	Characteristics of respondents included in the analysis are described	□	□
Modelling			
15	The dependent variable for each model is stated	□	□
16	Independent variables for each model are explained	□	□
17	Model specifications are provided	□	□
18	Model estimators are described	□	□
19	Goodness-of-fit statistics for each model are reported	□	□
Scoring algorithm			
20	Criteria for selecting the preferred model are stated	□	□
21	The scoring algorithm is presented	□	□

8.5. Appendix 5: Paper identification process for structured review

Figure 38: Structured review paper identification process



* Full text papers excluded for reporting the valuation of partial health states; valuing states not derived from a PBM; reporting qualitative work on a small amount of states; including DCE but not reporting the results in the paper.

8.6. Appendix 6: PRISMA Checklist

Table 66: PRISMA checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	2/4 (Identified as structured review)
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2 (review was not registered)
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-4
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	No interventions compared. Reference to questions on pg. 4
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	4
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4-5
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix 3
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	5-6
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	6
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	N/A (review of methods)
Summary	13	State the principal summary measures (e.g., risk ratio, difference	N/A

measures		in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	12
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Appendix 5
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Table 1
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarise the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	9-10
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	12
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	9-12
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	12

8.7. Appendix 7: Health Measurement Study - Survey Outline

Health Measurement Survey

Thank you for your interest in the health measurement survey! The survey is been carried out by the University of Technology Sydney!

In this study we are interested in the health of the Australian public. We are collecting this data from about 800 people to understand the health of the population in general. The data will also help us to improve the ways in which this information is collected in studies that are used to assess the effectiveness of new treatments for a range of health conditions.

If you agree to participate, we will ask you to complete a series of questions about your own health, including physical and mental health and wellbeing. Some of the questions may seem repetitive, but this helps us to understand the best ways to ask about your health. We would expect the survey to take about 20 minutes.

Your participation in this study is completely voluntary. You are not obliged to participate and may stop at any time. Your responses to the survey are strictly confidential and at no time will the answers you give be linked to your identity. The survey has been ethically approved by the University of Technology Sydney (App no: UTS HREC REF NO. 2015000135).

If you agree to take part, please click on the continue button below.

Demographics and self-reported health

Please indicate your age group

- 18 – 29 ☐
- 30 – 44 ☐
- 45 – 59 ☐
- 60 – 74 ☐
- 75+ ☐

What is your gender?

- Male ☐
- Female ☐

In general, would you say your health is?

- Excellent ☐
- Very good ☐
- Good ☐
- Fair ☐
- Poor ☐

How satisfied are you with your health?

Not at all
- 0 ☐
- 1 ☐
- 2 ☐
- 3 ☐
- 4 ☐
- 5 ☐
- 6 ☐
- 7 ☐
- 8 ☐
- 9 ☐
- 10 ☐

Completely

Do you have any illness, health problem, condition or disability?

- Yes ☐
- No ☐

If yes, please tick all that apply:

Tiredness/fatigue		High blood pressure	
Pain		Heart disease	
Insomnia		Osteoarthritis	
Anxiety/nerves		Stroke	
Depression		Cancer	
Diabetes		Other	
Breathing problems (e.g. asthma, emphysema)			

EQ-5D-5L

By placing a tick in one box in each group below, please indicate which statements best describe your own health TODAY.

MOBILITY

I have no problems in walking about ☐
I have slight problems in walking about ☐
I have moderate problems in walking about ☐
I have severe problems in walking about ☐
I am unable to walk about ☐

SELF-CARE

I have no problems washing or dressing myself ☐
I have slight problems washing or dressing myself ☐
I have moderate problems washing or dressing myself ☐
I have severe problems washing or dressing myself ☐
I am unable to wash or dress myself ☐

USUAL ACTIVITIES *(e.g. work, study, housework, family or leisure activities)*

I have no problems doing my usual activities ☐
I have slight problems doing my usual activities ☐
I have moderate problems doing my usual activities ☐
I have severe problems doing my usual activities ☐
I am unable to do my usual activities ☐

PAIN/DISCOMFORT

I have no pain or discomfort ☐
I have slight pain or discomfort ☐
I have moderate pain or discomfort ☐
I have severe pain or discomfort ☐
I have extreme pain or discomfort ☐

ANXIETY/DEPRESSION

I am not anxious or depressed ☐
I am slightly anxious or depressed ☐
I am moderately anxious or depressed ☐
I am severely anxious or depressed ☐

I am extremely anxious or depressed ☐

ASCOT

In this survey, we will describe social care related quality of life in a particular way, using a number of different areas such as control, safety and independence. To familiarise yourself with this approach, please answer these questions.

Which of the following statements best describes how much control you have over your daily life?

By 'control over daily life' we mean having the choice to do things or have things done as you like and when you want

I have as much control over my daily life as I want ☐

I have adequate control over my daily life ☐

I have some control over my daily life, but not enough ☐

I have no control over my daily life ☐

Thinking about keeping clean and presentable in appearance, which of the following statements best describes your situation?

I feel clean and are able to present yourself the way I like ☐

I feel adequately clean and presentable ☐

I feel less than adequately clean or presentable ☐

I don't feel at all clean or presentable ☐

Thinking about the food and drink you get, which of the following statements best describes your situation?

I get all the food and drink I like when I want ☐

I get adequate food and drink at OK times ☐

I don't always get adequate or timely food and drink ☐

I don't always get adequate or timely food and drink, and I think there is a risk to my health ☐

Which of the following statements best describes how safe you feel?

By 'feeling safe' we mean how safe you feel both inside and outside the home. This includes fear of abuse, falling or other physical harm

I feel as safe as I want ☐

Generally, I feel adequately safe, but not as safe as I would like ☐

I feel less than adequately safe ☐

I don't feel at all safe ☐

Thinking about how much contact you have with people you like, which of the following statements best describes your social situation?

I have as much social contact as I want with people I like ☐

I have adequate social contact with people ☐

I have some social contact with people, but not enough ☐

I have little social contact with people and feel socially isolated ☐

Which of the following statements best describes how you spend your time?

- I'm able to spend time as I want, doing things I value or enjoy ☐
- I'm able to do enough of the things I value or enjoy with my time ☐
- I do some of the things I value or enjoy with my time, but not enough ☐
- I don't do anything I value or enjoy with my time ☐

Which of the following statements best describes how clean and comfortable your home is?

- My home is as clean and comfortable as I want ☐
- My home is adequately clean and comfortable ☐
- My home is not quite clean or comfortable enough ☐
- My home is not at all clean or comfortable ☐

Which of these statements best describes how having help to do things makes you think and feel about yourself?

- Having help makes me think and feel better about myself ☐
- Having help does not affect the way I think or feel about myself ☐
- Having help sometimes undermines the way I think and feel about myself ☐
- Having help completely undermines the way I think and feel about myself ☐

Which of these statements best describes how the way you are helped and treated makes you think and feel about yourself?

- The way I'm helped and treated makes me think and feel better about myself ☐
- The way I'm helped and treated does not affect the way I think or feel about myself ☐
- The way I'm helped and treated sometimes undermines the way I think and feel about myself ☐
- The way I'm helped and treated completely undermines the way I think and feel about myself ☐

ICECAP-A

Please indicate which statements best describe your overall quality of life at the moment by placing a tick in ONE box for each of the five groups below.

Feeling settled and secure

- I am able to feel settled and secure in all areas of my life ☐
- I am able to feel settled and secure in many areas of my life ☐
- I am able to feel settled and secure in a few areas of my life ☐
- I am unable to feel settled and secure in any areas of my life ☐

Love, friendship and support

- I can have a lot of love, friendship and support ☐
- I can have quite a lot of love, friendship and support ☐
- I can have a little love, friendship and support ☐
- I cannot have any love, friendship and support ☐

Being independent

- I am able to be completely independent ☐
- I am able to be independent in many things ☐
- I am able to be independent in a few things ☐
- I am unable to be at all independent ☐

Achievement and progress

- I can achieve and progress in all aspects of my life ☐
- I can achieve and progress in many aspects of my life ☐
- I can achieve and progress in a few aspects of my life ☐
- I cannot achieve and progress in any aspects of my life ☐

Enjoyment and pleasure

- I can have a lot of enjoyment and pleasure ☐
- I can have quite a lot of enjoyment and pleasure ☐
- I can have a little enjoyment and pleasure ☐
- I cannot have any enjoyment and pleasure ☐

SF-36v2

1. In general would you say your health is:

Excellent	☐
Very good	☐
Good	☐
Fair	☐
Poor	☐

2. Compared to one year ago, how would you rate your health in general now?

Much better now than one year ago	☐
Somewhat better now than one year ago	☐
About the same as one year ago	☐
Somewhat worse now than one year ago	☐
Much worse now than one year ago	☐

3. The following questions are about activities you might do during a typical week day. Does your health now limit you in these activities? If so how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
a. Vigourous activities, such as running, lifting heavy objects, participating in strenuous sports	☐	☐	☐
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	☐	☐	☐
c. Lifting or carrying groceries	☐	☐	☐
d. Climbing several flights of stairs	☐	☐	☐
e. Climbing one flight of stairs	☐	☐	☐
f. Bending, kneeling, or stooping	☐	☐	☐
g. Walking more than one kilometre	☐	☐	☐
h. Walking half a kilometre	☐	☐	☐
i. Walking 100 metres	☐	☐	☐
j. Bathing or dressing yourself	☐	☐	☐

4. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a. Cut down on the amount of time you spent on work and other activities					
b. Accomplished less than you would like					
c. Were limited in the kind of work or other activities					
d. Had difficulty performing the work or other activities (for example it took extra effort)					

5. During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a. Cut down on the amount of time you spent on work and other activities					
b. Accomplished less than you would like					
c. Did work or other activities less carefully than usual					

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours or groups?

Not at all	
Slightly	
Moderately	
Quite a bit	
Extremely	

7. How much bodily pain have you had during the past 4 weeks?

None	
Very mild	
Mild	
Moderate	
Severe	
Very severe	

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

Not at all	☐
A little bit	☐
Moderately	☐
Quite a bit	☐
Extremely	☐

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks…

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
Did you feel full of life?	☐	☐	☐	☐	☐
Have you been very nervous?	☐	☐	☐	☐	☐
Have you felt so down in the dumps that nothing could cheer you up?	☐	☐	☐	☐	☐
Have you felt calm and peaceful?	☐	☐	☐	☐	☐
Did you have a lot of energy?	☐	☐	☐	☐	☐
Have you felt downhearted and depressed?	☐	☐	☐	☐	☐
Did you feel worn out?	☐	☐	☐	☐	☐
Have you been happy?	☐	☐	☐	☐	☐
Did you feel tired?	☐	☐	☐	☐	☐

10. During the past 4 weeks how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?

All of the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

11. How TRUE or FALSE is each of the following statements for you?

	Definitely true	Mostly true	Don't know	Mostly false	Definitely false
I seem to get sick a little easier than other people	☐	☐	☐	☐	☐
I am as healthy as anybody I know	☐	☐	☐	☐	☐
I expect my health to get worse	☐	☐	☐	☐	☐
My health is excellent	☐	☐	☐	☐	☐

PROMIS–29v2

Please respond to each question or statement by marking one box per row.

Physical Function

1 Are you able to do chores such as vacuuming or yard work?

- Without any difficulty ☐
- With a little difficulty ☐
- With some difficulty ☐
- With much difficulty ☐
- Unable to do ☐

2 Are you able to go up and down stairs at a normal pace?

- Without any difficulty ☐
- With a little difficulty ☐
- With some difficulty ☐
- With much difficulty ☐
- Unable to do ☐

3 Are you able to go for a walk of at least 15 minutes?

- Without any difficulty ☐
- With a little difficulty ☐
- With some difficulty ☐
- With much difficulty ☐
- Unable to do ☐

4 Are you able to run errands and shop?

- Without any difficulty ☐
- With a little difficulty ☐
- With some difficulty ☐
- With much difficulty ☐
- Unable to do ☐

5 In the past 7 days I felt fearful

- Never ☐
- Rarely ☐
- Sometimes ☐
- Often ☐
- Always ☐

6 In the past 7 days I found it hard to focus on anything other than my anxiety

- Never ☐
- Rarely ☐
- Sometimes ☐
- Often ☐

Always ☐

7 In the past 7 days my worries overwhelmed me
Never ☐
Rarely ☐
Sometimes ☐
Often ☐
Always ☐

8 In the past 7 days I felt uneasy
Never ☐
Rarely ☐
Sometimes ☐
Often ☐
Always ☐

9 In the past 7 days I felt worthless
Never ☐
Rarely ☐
Sometimes ☐
Often ☐
Always ☐

10 In the past 7 day I felt helpless
Never ☐
Rarely ☐
Sometimes ☐
Often ☐
Always ☐

11 In the past 7 days I felt depressed
Never ☐
Rarely ☐
Sometimes ☐
Often ☐
Always ☐

12 In the past 7 days I felt hopeless
Never ☐
Rarely ☐
Sometimes ☐
Often ☐
Always ☐

13 During the past 7 days I feel fatigued

- Not at all ☐
- A little bit ☐
- Somewhat ☐
- Quite a bit ☐
- Very much ☐

14 During the past 7 days I have trouble starting things because I am tired

- Not at all ☐
- A little bit ☐
- Somewhat ☐
- Quite a bit ☐
- Very much ☐

15 In the past 7 days how run-down did you feel on average?...

- Not at all ☐
- A little bit ☐
- Somewhat ☐
- Quite a bit ☐
- Very much ☐

16 In the past 7 days how fatigued were you on average?

- Not at all ☐
- A little bit ☐
- Somewhat ☐
- Quite a bit ☐
- Very much ☐

17 In the past 7 days my sleep quality was

- Very poor ☐
- Poor ☐
- Fair ☐
- Good ☐
- Very good ☐

18 In the past 7 days my sleep was refreshing

- Not at all ☐
- A little bit ☐
- Somewhat ☐
- Quite a bit ☐
- Very much ☐

19 In the past 7 days I had a problem with my sleep

- Not at all ☐
- A little bit ☐
- Somewhat ☐
- Quite a bit ☐
- Very much ☐

20 In the past 7 days I had difficulty falling asleep

- Not at all ☐
- A little bit ☐
- Somewhat ☐
- Quite a bit ☐
- Very much ☐

21 I have trouble doing all of my regular leisure activities with others

- Never ☐
- Rarely ☐
- Sometimes ☐
- Often ☐
- Always ☐

22 I have trouble doing all of the family activities that I want to do

- Never ☐
- Rarely ☐
- Sometimes ☐
- Often ☐
- Always ☐

23 I have trouble doing all of my usual work (include work at home)

- Never ☐
- Rarely ☐
- Sometimes ☐
- Often ☐
- Always ☐

24 I have trouble doing all of the activities with friends that I want to do

- Never ☐
- Rarely ☐
- Sometimes ☐
- Often ☐
- Always ☐

25 In the past 7 days how much did pain interfere with your day to day activities?

Not at all ☐
A little bit ☐
Somewhat ☐
Quite a bit ☐
Very much ☐

26 In the past 7 days how much did pain interfere with work around the home?
Not at all ☐
A little bit ☐
Somewhat ☐
Quite a bit ☐
Very much ☐

27 In the past 7 days how much did pain interfere with your ability to participate in social activities?
Not at all ☐
A little bit ☐
Somewhat ☐
Quite a bit ☐
Very much ☐

28 In the past 7 days how much did pain interfere with your household chores?
Not at all ☐
A little bit ☐
Somewhat ☐
Quite a bit ☐
Very much ☐

29 In the past 7 days how would you rate your pain on average?
No pain
0 ☐
1 ☐
2 ☐
3 ☐
4 ☐
5 ☐
6 ☐
7 ☐
8 ☐
9 ☐
10 ☐
Worst imaginable pain

WEMWBS

Below are some statements about feelings and thoughts. Please tick the box that best describes your experience over the past 2 weeks

I've been feeling optimistic about the future

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐
- All of the time ☐

I've been feeling useful

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐
- All of the time ☐

I've been feeling relaxed

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐
- All of the time ☐

I've been feeling interested in other people

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐
- All of the time ☐

I've had energy to spare

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐
- All of the time ☐

I've been dealing with problems well

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐

All of the time ☐

I've been thinking clearly
None of the time ☐
Rarely ☐
Some of the time ☐
Often ☐
All of the time ☐

I've been feeling good about myself
None of the time ☐
Rarely ☐
Some of the time ☐
Often ☐
All of the time ☐

I've been feeling close to other people
None of the time ☐
Rarely ☐
Some of the time ☐
Often ☐
All of the time ☐

I've been feeling confident
None of the time ☐
Rarely ☐
Some of the time ☐
Often ☐
All of the time ☐

I've been able to make up my own mind about things
None of the time ☐
Rarely ☐
Some of the time ☐
Often ☐
All of the time ☐

I've been feeling loved
None of the time ☐
Rarely ☐
Some of the time ☐
Often ☐
All of the time ☐

I've been interested in new things

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐
- All of the time ☐

I've been feeling cheerful

- None of the time ☐
- Rarely ☐
- Some of the time ☐
- Often ☐
- All of the time ☐

ONS-4

Overall, how satisfied are you with your life nowadays?

Not at all

- 0 ☐
- 1 ☐
- 2 ☐
- 3 ☐
- 4 ☐
- 5 ☐
- 6 ☐
- 7 ☐
- 8 ☐
- 9 ☐
- 10 ☐

Completely

Overall, to what extent do you feel the things you do in your life are worthwhile?

Not at all

- 0 ☐
- 1 ☐
- 2 ☐
- 3 ☐
- 4 ☐
- 5 ☐
- 6 ☐
- 7 ☐
- 8 ☐

9 ☐
10 ☐
Completely

Overall, how happy were you yesterday?
Not at all
0 ☐
1 ☐
2 ☐
3 ☐
4 ☐
5 ☐
6 ☐
7 ☐
8 ☐
9 ☐
10 ☐
Completely

Overall, how anxious were you feeling yesterday?
Not at all
0 ☐
1 ☐
2 ☐
3 ☐
4 ☐
5 ☐
6 ☐
7 ☐
8 ☐
9 ☐
10 ☐
Completely

FURTHER DEMOGRAPHICS
We now need to collect some data about you.
What is your country of birth?
Australia ☐
United Kingdom ☐
New Zealand ☐
Italy ☐
Vietnam ☐

Greece	☐
Germany	☐
Philippines	☐
China	☐
Indonesia	☐
Other	☐

If other, please specify_____

What is the highest level of education you completed?

Primary	☐
Secondary	☐
Trade certificate/Diploma	☐
Bachelor's degree	☐
Higher degree	☐

Are you currently studying?

Yes	☐
No	☐

Please indicate which of these categories best matches your gross (before tax) income?

- Under $20,000 ☐
- $20,001-$30,000 ☐
- $30,001-$40,000 ☐
- $40,001-$50,000 ☐
- $50,001-$60,000 ☐
- $60,001-$70,000 ☐
- $70,001-$80,000 ☐
- $80,001-$100,000 ☐
- Over $100,000 ☐
- Prefer not to say ☐

What is your current marital status?

- Single ☐
- Separated/divorced ☐
- Widowed ☐
- Married/De facto ☐

How many children or dependents do you have?

- None ☐
- 1 ☐
- 2 ☐
- 3 or more ☐

Approximately how many times have you seen a family doctor or another GP about your health in the last 12 months?
TEXT BOX

On how many different occasions were you admitted as a patient to a hospital for an overnight stay during the last 12 months?
TEXT BOX

8.8. Appendix 8: Orthogonal EFA models

Table 67 and **Table 68** report the dimension structure for the orthogonal CF-Varimax and CF-Quartimax models respectively. The blanks indicate factor loadings below the minimum level of 0.3.

Table 67: EFA Orthogonal CF-Varimax model (Model 3)

	Dim 1[a]	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10
Item	**FL (SE)[b]**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**
EQ MO			0.88 (0.00)							
EQ SC	0.34 (0.04)		0.78 (0.03)							0.35 (0.02)
EQ UA	0.32 (0.03)		0.74 (0.02)							
EQ PD			0.68 (0.03)						0.39 (0.01)	
EQ AD	0.07 (0.02)									
ASCOT 1	0.38 (0.04)	0.48 (0.03)								
ASCOT 2	0.51 (0.04)	0.34 (0.04)								0.35 (0.02)
ASCOT 3	0.43 (0.05)	0.35 (0.04)								0.36 (0.04)
ASCOT 4	0.44 (0.05)	0.43 (0.04)	0.36 (0.05)							
ASCOT 5	0.43 (0.04)	0.56 (0.03)						0.32 (0.04)		
ASCOT 6	0.37 (0.04)	0.55 (0.04)				0.32 (0.04)		0.30 (0.04)		
ASCOT 7	0.32 (0.05)	0.36 (0.05)						0.31 (0.06)		
ASCOT 8		0.33 (0.05)			0.37 (N/R)					
ASCOT 9		0.34 (0.06)			0.31 (0.08)					
ICECAP 1	0.48 (0.04)	0.50 (0.04)						0.39 (0.04)		
ICECAP 2	0.40 (0.04)	0.56 (0.05)						0.43 (0.04)		
ICECAP 3	0.42 (0.05)	0.41 (0.05)	0.37 (0.05)							
ICECAP 4	0.41 (0.04)	0.54 (0.06)						0.42 (0.04)		
ICECAP 5	0.46 (0.04)	0.53 (0.04)						0.41 (0.04)		
SF-36 3			0.78 (0.03)							
SF-36 4			0.84 (0.02)							
SF-36 5			0.83 (0.02)							
SF-36 6			0.84 (0.02)							
SF-36 7			0.89 (0.02)							
SF-36 8			0.82 (0.02)							
SF-36 9			0.90 (0.01)							
SF-36 10			0.91 (0.01)							
SF-36 11			0.88 (0.02)							

SF-36 12	0.33 (0.06)	0.81 (0.03)					
SF-36 13	0.32 (0.03)	0.74 (0.03)			0.35 (0.05)		
SF-36 14	0.39 (0.03)	0.59 (0.03)			0.48 (0.07)		
SF-36 15	0.31 (0.03)	0.77 (0.02)			0.39 (0.03)		
SF-36 16	0.34 (0.03)	0.74 (0.02)			0.38 (0.04)		
SF-36 17	0.50 (0.03)	0.64 (0.02)			0.42 (0.03)		
SF-36 18	0.54 (0.03)	0.53 (0.03)			0.41 (0.03)		
SF-36 19	0.58 (0.03)	0.49 (0.03)			0.35 (0.04)		
SF-36 20	0.60 (0.03)	0.47 (0.03)					
SF-36 21		0.68 (0.02)					0.44 (0.03)
SF-36 22		0.70 (0.02)					0.40 (0.03)
SF-36 23	0.36 (0.04)			0.34 (0.03)			
SF-36 24	0.76 (0.03)						
SF-36 25	0.79 (0.02)						
SF-36 26	0.53 (0.03)					0.52 (0.03)	
SF-36 27						0.56 (0.04)	
SF-36 28	0.78 (0.02)						
SF-36 29	0.45 (0.03)	0.39 (0.03)		0.52 (0.03)			
SF-36 30	0.53 (0.03)					0.59 (0.03)	
SF-36 31	0.42 (0.04)	0.32 (0.04)		0.60 (0.03)			
SF-36 32	0.60 (0.03)	0.47 (0.03)					
PROMIS 1		0.79 (0.02)					
PROMIS 2		0.82 (0.02)					
PROMIS 3		0.87 (0.02)					
PROMIS 4	0.31 (0.04)	0.80 (0.02)					
PROMIS 5	0.75 (0.02)	0.33 (0.04)					
PROMIS 6	0.83 (0.02)						
PROMIS 7	0.81 (0.02)						
PROMIS 8	0.76 (0.02)	0.30 (0.03)					
PROMIS 9	0.77 (0.02)					0.34 (0.03)	
PROMIS 10	0.76 (0.02)	0.36 (0.03)					
PROMIS 11	0.80 (0.02)					0.32 (0.02)	
PROMIS 12	0.78 (0.02)	0.31 (0.03)				0.30 (0.02)	
PROMIS 13	0.42 (0.02)	0.34 (0.03)		0.74 (0.02)			
PROMIS 14	0.40 (N/R)	0.30 (0.02)		0.73 (0.02)			
PROMIS 15	0.50 (0.02)	0.35 (0.03)		0.67 (0.02)			
PROMIS 16	0.43 (0.03)	0.34 (0.04)		0.74 (0.02)			
PROMIS 17	0.37 (0.03)		0.46 (0.03)			0.42 (0.03)	
PROMIS 18	0.32 (0.03)		0.42 (0.03)	0.35 (0.02)		0.55 (0.03)	
PROMIS 19	0.48 (0.03)	0.30 (0.04)	0.50 (0.03)	0.31 (0.03)			
PROMIS 20	0.55 (0.03)		0.39 (0.04)				
PROMIS 21	0.49 (0.03)	0.61 (0.02)					
PROMIS 22	0.44 (0.02)	0.62 (0.02)					
PROMIS 23	0.39 (0.03)	0.66 (0.02)					

PROMIS 24	0.48 (0.03)	0.60 (0.02)			
PROMIS 25	0.31 (0.03)	0.73 (0.02)			0.47 (0.04)
PROMIS 26	0.30 (0.03)	0.75 (0.01)			0.46 (0.02)
PROMIS 27	0.37 (0.03)	0.71 (0.02)			0.38 (0.03)
PROMIS 28		0.76 (0.01)			0.43 (0.02)
WEMWBS 1	0.48 (0.03)			0.60 (0.03)	
WEMWBS 2	0.54 (0.03)	0.32 (0.03)		0.57 (0.03)	
WEMWBS 3	0.56 (0.03)			0.54 (0.03)	
WEMWBS 4	0.41 (0.04)			0.60 (0.03)	
WEMWBS 5		0.35 (0.04)		0.59 (0.03)	
WEMWBS 6	0.61 (0.03)			0.51 (0.03)	
WEMWBS 7	0.64 (0.03)			0.45 (0.03)	
WEMWBS 8	0.56 (0.02)			0.65 (0.02)	
WEMWBS 9	0.56 (0.02)			0.64 (0.03)	
WEMWBS 10	0.58 (0.02)			0.62 (0.02)	
WEMWBS 11	0.67 (0.03)			0.41 (0.03)	
WEMWBS 12	0.48 (0.03)			0.54 (0.04)	
WEMWBS 13	0.42 (0.04)			0.57 (0.03)	
WEMWBS 14	0.57 (0.03)			0.65 (0.02)	

[a] Dimension; [b] Factor loading (standard error); Note: fuller item descriptions in Appendix 9

Table 68: EFA Orthogonal quartimax model (Model 4)

	Dim 1[a]	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10
Item	**FL (SE)[b]**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**	**FL (SE)**
EQ MO	0.74 (0.01)				0.52 (0.03)					
EQ SC	0.60 (0.04)				0.58 (0.05)			0.30 (0.02)		
EQ UA	0.53 (0.03)				0.70 (0.03)					
EQ PD	0.51 (0.03)			0.38 (0.02)	0.57 (0.03)					
EQ AD					0.84 (0.02)					
ASCOT 1					0.67 (0.04)	0.30 (0.04)				
ASCOT 2					0.74 (0.04)					
ASCOT 3					0.61 (0.05)					
ASCOT 4					0.64 (0.05)					
ASCOT 5					0.70 (0.03)	0.42 (0.04)				
ASCOT 6					0.70 (0.03)	0.41 (0.04)				
ASCOT 7					0.57 (0.05)					
ASCOT 8			0.33 (N/R)			0.33 (0.05)				
ASCOT 9					0.35 (0.07)	0.30 (0.06)				
ICECAP 1					0.77 (0.03)	0.35 (0.04)				
ICECAP 2					0.69 (0.04)	0.45 (0.05)				
ICECAP 3					0.66 (0.04)					
ICECAP 4					0.76 (0.03)	0.39 (0.07)				
ICECAP 5					0.78 (0.03)	0.37 (0.05)				
SF-36 3	0.72 (0.03)				0.33 (0.05)					
SF-36 4	0.71 (0.03)				0.50 (0.04)					
SF-36 5	0.71 (0.03)				0.48 (0.05)					
SF-36 6	0.75 (0.03)				0.44 (0.04)					
SF-36 7	0.77 (0.03)				0.49 (0.04)					
SF-36 8	0.71 (0.03)				0.46 (0.04)					
SF-36 9	0.79 (0.02)				0.48 (0.03)					
SF-36 10	0.80 (0.02)				0.49 (0.04)					
SF-36 11	0.74 (0.03)				0.53 (0.04)					
SF-36 12	0.65 (0.05)				0.54 (0.06)					
SF-36 13	0.55 (0.03)				0.60 (0.03)		0.33 (0.05)			
SF-36 14	0.37 (0.03)				0.68 (0.03)		0.45 (0.08)			
SF-36 15	0.58 (0.02)				0.63 (0.03)		0.37 (0.03)			
SF-36 16	0.54 (0.02)				0.68 (0.02)		0.36 (0.04)			
SF-36 17	0.41 (0.03)				0.70 (0.02)		0.39 (0.03)			
SF-36 18					0.77 (0.02)		0.37 (0.04)			
SF-36 19					0.70 (0.03)		0.31 (0.04)			
SF-36 20					0.81 (0.02)					

SF-36 21	0.50 (0.03)		0.43 (0.03)		0.58 (0.03)		
SF-36 22	0.50 (0.03)		0.37 (0.03)		0.65 (0.03)		
SF-36 23					0.75 (0.03)		
SF-36 24					0.76 (0.03)		
SF-36 25					0.86 (0.02)		
SF-36 26					0.77 (0.03)		
SF-36 27					0.69 (0.04)	0.48 (0.06)	
SF-36 28					0.88 (0.01)		
SF-36 29					0.77 (0.02)		
SF-36 30					0.80 (0.02)		
SF-36 31					0.74 (0.03)		0.41 (0.03)
SF-36 32					0.81 (0.02)		
PROMIS 1	0.62 (0.03)				0.60 (0.03)		
PROMIS 2	0.70 (0.03)				0.51 (0.04)		
PROMIS 3	0.72 (0.02)				0.57 (0.03)		
PROMIS 4	0.61 (0.03)				0.65 (0.03)		
PROMIS 5					0.81 (0.02)		
PROMIS 6					0.87 (0.02)		
PROMIS 7					0.87 (0.02)		
PROMIS 8					0.86 (0.02)		
PROMIS 9					0.92 (0.01)		
PROMIS 10					0.92 (0.01)		
PROMIS 11					0.92 (0.01)		
PROMIS 12					0.93 (0.01)		
PROMIS 13					0.79 (0.02)		0.53 (0.03)
PROMIS 14					0.74 (0.02)		0.54 (0.02)
PROMIS 15					0.83 (0.01)		0.44 (0.02)
PROMIS 16					0.80 (0.01)		0.53 (0.02)
PROMIS 17		0.41 (0.03)			0.71 (0.03)		
PROMIS 18		0.36 (0.03)			0.71 (0.02)		
PROMIS 19		0.44 (0.03)			0.72 (0.03)		
PROMIS 20		0.34 (0.04)			0.71 (0.03)		
PROMIS 21	0.35 (0.03)				0.80 (0.02)		
PROMIS 22	0.37 (0.02)				0.79 (0.02)		
PROMIS 23	0.42 (0.03)				0.77 (0.02)		
PROMIS 24	0.34 (0.03)				0.81 (0.02)		
PROMIS 25	0.51 (0.02)			0.45 (0.04)	0.68 (0.03)		
PROMIS 26	0.53 (0.02)			0.43 (0.02)	0.69 (0.02)		

PROMIS 27	0.47 (0.03)	0.35 (0.03)	0.72 (0.02)	
PROMIS 28	0.55 (0.02)	0.40 (0.02)	0.69 (0.02)	
WEMWBS 1			0.73 (0.03)	
WEMWBS 2			0.82 (0.03)	
WEMWBS 3			0.84 (0.02)	
WEMWBS 4			0.68 (0.03)	
WEMWBS 5			0.70 (0.03)	
WEMWBS 6			0.81 (0.02)	
WEMWBS 7			0.80 (0.02)	
WEMWBS 8			0.88 (0.01)	
WEMWBS 9			0.76 (0.02)	0.36 (0.03)
WEMWBS 10			0.87 (0.01)	
WEMWBS 11			0.77 (0.02)	
WEMWBS 12			0.72 (0.03)	0.31 (0.04)
WEMWBS 13			0.72 (0.03)	
WEMWBS 14			0.85 (0.02)	

[a] Dimension; [b] Factor loading (standard error); Note: fuller item descriptions in Appendix 9

8.9. Appendix 9: Item coding and further item description for the measurement chapter

Table 69: Item descriptions for the 91 items included in the IRT analyses

Coding used in results tables	Further item description
EQ MO	Problems with walking around
EQ SC	Problems with washing or dressing
EQ UA	Problems doing usual activities
EQ PD	Pain or discomfort
EQ AD	Anxiety or depression
ASCOT 1	Control over daily life
ASCOT 2	Keeping clean and presentable in appearance
ASCOT 3	Thinking about the food and drink you get
ASCOT 4	How safe you feel
ASCOT 5	How much contact you have with people you like
ASCOT 6	How you spend your time
ASCOT 7	How clean and comfortable home is
ASCOT 8	How having help to do things makes you think and feel
ASCOT 9	How the way you are helped and treated makes you think and feel
ICECAP 1	Feeling settled and secure
ICECAP 2	Love, friendship and support
ICECAP 3	Being independent
ICECAP 4	Achievement and progress
ICECAP 5	Enjoyment and pleasure
SF-36 3	Vigorous activity limitations
SF-36 4	Moderate activity limitations
SF-36 5	Lifting or carrying groceries
SF-36 6	Climbing several flights of stairs
SF-36 7	Climbing one flight of stairs
SF-36 8	Bending, kneeling or stooping
SF-36 9	Walking more than one kilometre
SF-36 10	Walking several hundred metres
SF-36 11	Walking one hundred metres
SF-36 12	Bathing or dressing yourself
SF-36 13	Cut down on the amount of time you spent on work or other activities as a result of physical health
SF-36 14	Accomplished less than you would like as a result of physical health
SF-36 15	Limited in the kind of work or other activities as a result of physical health
SF-36 16	difficulty performing the work or other activities as a result of physical health
SF-36 17	Cut down on the amount of time you spent on work or other activities as a result of emotional problems
SF-36 18	Accomplished less than you would like as a result of emotional problems
SF-36 19	Did work or other activities less carefully than usual as a result of emotional problems
SF-36 20	physical health or emotional problems interfere with normal social activities
SF-36 21	How much bodily pain
SF-36 22	How much pain interfered with normal work
SF-36 23	Feel full of life
SF-36 24	Been very nervous
SF-36 25	So down in the dumps that nothing could cheer you up
SF-36 26	Felt calm and peaceful

SF-36 27	Have a lot of energy
SF-36 28	Felt downhearted and depressed
SF-36 29	Feel worn out
SF-36 30	Been happy
SF-36 31	Feel tired
SF-36 32	Physical health or emotional problems interfered with social activities
PROMIS 1	Able to do chores such as vacuuming or yard work
PROMIS 2	Able to go up and down stairs at a normal pace
PROMIS 3	Able to go for a walk of at least 15 minutes
PROMIS 4	Able to run errands and shop
PROMIS 5	Felt fearful
PROMIS 6	Found it hard to focus on anything other than my anxiety
PROMIS 7	Worries overwhelming
PROMIS 8	Felt uneasy
PROMIS 9	Felt worthless
PROMIS 10	Felt helpless
PROMIS 11	Felt depressed
PROMIS 12	Felt hopeless
PROMIS 13	Feel fatigued
PROMIS 14	Have trouble starting things because tired
PROMIS 15	How run-down
PROMIS 16	How fatigued
PROMIS 17	Sleep Quality
PROMIS 18	Sleep was refreshing
PROMIS 19	Problem with sleep
PROMIS 20	Difficulty falling asleep
PROMIS 21	Trouble doing all regular leisure activities with others
PROMIS 22	Trouble doing all family activities
PROMIS 23	Trouble doing usual work (include work at home)
PROMIS 24	Trouble doing all activities with friends
PROMIS 25	How much did pain interfere with day to day activities
PROMIS 26	How much did pain interfere with work around the home
PROMIS 27	How much did pain interfere with ability to participate in social activities
PROMIS 28	How much did pain interfere with household chores
WEMWBS 1	Feeling optimistic about the future
WEMWBS 2	Feeling useful
WEMWBS 3	Feeling relaxed
WEMWBS 4	Feeling interested in other people
WEMWBS 5	Energy to spare
WEMWBS 6	Dealing with problems well
WEMWBS 7	Thinking clearly
WEMWBS 8	Feeling good about self
WEMWBS 9	Feeling close to other people
WEMWBS 10	Feeling confident
WEMWBS 11	Able to make up mind about things
WEMWBS 12	Feeling loved
WEMWBS 13	Interested in new things
WEMWBS 14	Feeling cheerful

8.10. Appendix 10: Local independence values (all item pairs)

Table 70 to Table 73 report the local dependence Chi Square values for the physical functioning, mental health, pain and activities dimensions

Table 70: Local dependencies Chi Square values across item pairs – Physical functioning dimension[a]

	EQ MO	EQ SC	EQ UA	SF-36 3	SF-36 4	SF-36 5	SF-36 6	SF-36 7	SF-36 8	SF-36 9	SF-36 10	SF-36 11	SF-36 12	PROMIS 1	PROMIS 2	PROMIS 3
EQ MO	-															
EQ SC	3.3	-														
EQ UA	5.0	4.8	-													
SF-36 3	2.0	8.7	0.1	-												
SF-36 4	3.7	3.9	2.9	**11.7**	-											
SF-36 5	2.5	1.8	5.1	2.1	**12.7**	-										
SF-36 6	2.0	5.3	3.0	5.3	5.5	0.4	-									
SF-36 7	1.7	1.1	6.8	4.2	5.0	5.6	7.9	-								
SF-36 8	0.8	2.8	2.6	4.0	1.1	0.1	0.6	0.9	-							
SF-36 9	1.2	6.7	2.0	1.3	4.4	0.3	0.3	0.4	1.0	-						
SF-36 10	3.2	3.4	1.9	1.4	3.0	0.2	0.3	0.5	0.2	**10.4**	-					
SF-36 11	4.3	0.3	3.1	5.7	3.9	1.0	1.2	0.3	0.2	0.1	5.8	-				
SF-36 12	6.4	**14.1**	4.5	6.4	1.5	4.1	1.7	1.3	0.7	3.9	6.1	6.1	-			
PROMIS 1	2.9	2.1	2.8	4.1	**14.5**	**12.9**	**14.9**	**12.3**	1.7	6.1	6.6	2.5	17.7	-		
PROMIS 2	1.1	0.8	1.8	1.0	**17.0**	**14.1**	2.3	**17.0**	1.9	8.7	4.7	3.8	12.0	7.4	-	
PROMIS 3	0.7	1.2	1.1	4.0	6.5	7.1	3.8	7.7	4.4	6.5	5.4	8.4	4.8	2.3	3.7	-
PROMIS 4	4.9	1.1	5.3	8.5	6.5	3.1	**11.8**	5.8	4.2	9.7	6.0	6.0	4.2	5.3	1.9	**11.7**

[a] Standardised Chi Square values > 10 could be indicative of local dependency (and are bolded)

Table 71: Local dependence estimates - Mental health dimension[a]

Item	EQ-5D-5L AD	SF-36 24	SF-36 25	SF-36 26	SF-36 28	SF-36 30	PROMIS 5	PROMIS 6	PROMIS 7	PROMIS 8	PROMIS 9	PROMIS 10
EQ-5D-5L AD	-											
SF-36 24	0.6	-										
SF-36 25	1.8	5.9	-									
SF-36 26	6.9	**13.3**	**18.7**	-								
SF-36 28	2.0	4.0	**11.5**	**17.6**	-							
SF-36 30	5.9	**12.0**	**18.8**	**25.7**	**28.1**	-						
PROMIS 5	0.7	7.4	6.2	9.7	7.4	**12.0**	-					
PROMIS 6	0.1	4.7	3.0	**13.4**	3.9	6.8	8.6	-				
PROMIS 7	5.6	4.2	1.9	7.9	8.7	**12.0**	**12.6**	6.6	-			
PROMIS 8	0.4	3.0	1.8	**22.9**	2.8	**15.7**	9.0	4.2	3.5	-		
PROMIS 9	4.7	5.7	1.2	9.0	1.8	8.1	**13.2**	**11.9**	7.0	1.6	-	
PROMIS 10	1.8	3.7	2.4	**11.5**	2.7	**10.5**	4.2	7.5	4.8	1.5	8.6	-
PROMIS 11	0.7	6.9	4.8	**10.7**	**12.6**	**11.9**	9.7	4.1	0.6	1.0	2.6	1.1

[a] Standardised Chi Square values > 10 could be indicative of local dependency (and are bolded)

Table 72: Local dependence estimates - Pain dimension[a]

Item	EQ-5D-5L PD	SF-36 21	SF-36 22	PROMIS 1	PROMIS 25	PROMIS 26	PROMIS 27
EQ-5D-5L PD	-						
SF-36 21	**18.2**	-					
SF-36 22	6.2	8.9	-				
PROMIS 1	5.3	1.2	1.4	-			
PROMIS 25	**12.6**	0.1	4.7	1.0	-		
PROMIS 26	**10.8**	1.8	1.3	3.5	6.8	-	
PROMIS 27	**11.0**	4.0	1.4	2.1	3.6	2.4	
PROMIS 28	**11.4**	4.4	0.6	4.2	5.9	7.0	4.3

[a] Standardised Chi Square values > 10 could be indicative of local dependency (and are bolded)

Table 73: Local independence - Activities dimension[a]

Item	EQ-5D-5L PD	SF-36 20	SF-36 32	PROMIS 21	PROMIS 22	PROMIS 23
EQ-5D-5L UA	-					
SF-36 20	2.2	-				
SF-36 32	1.4	**25.8**	-			
PROMIS 21	1.5	3.6	5.1	-		
PROMIS 22	1.5	2.2	2.5	6.0	-	
PROMIS 23	0.7	4.6	3.8	3.9	6.8	-
PROMIS 24	2.3	4.3	2.3	6.9	7.4	**11.4**

8.11. Appendix 11: One Block of choice sets from the DCE design

Table 74: Block of choice sets from the Chapter 5 DCE design

Pair no	Pair code	MO	SC	UA	PD	AD	CO	CL	FD	SA	SP	AC	OC	DI
5	A	5	2	3	1	3	4	2	3	1	3	1	2	4
5	B	3	4	2	5	3	4	1	3	1	4	2	2	2
6	A	3	3	5	3	4	1	4	4	3	2	3	4	3
6	B	4	3	2	3	3	4	2	2	4	2	2	4	3
50	A	5	5	5	5	3	1	2	3	3	4	3	1	2
50	B	5	4	1	1	3	1	4	3	4	3	3	2	1
91	A	4	2	1	4	5	1	2	1	4	1	2	4	3
91	B	4	1	5	4	1	4	2	2	2	1	4	4	1
146	A	1	2	3	2	4	2	3	4	2	4	1	3	1
146	B	1	4	1	3	1	2	3	1	1	4	4	1	1
173	A	2	2	5	5	5	3	4	2	2	4	3	4	1
173	B	1	3	5	5	4	4	1	3	4	4	3	2	1
184	A	1	3	1	4	5	2	2	3	1	1	3	1	4
184	B	1	1	3	4	1	4	2	4	2	2	2	1	4
201	A	1	3	2	3	2	3	4	2	2	2	2	2	3
201	B	5	3	1	5	2	4	3	4	2	3	2	2	1
206	A	1	3	5	5	1	1	1	3	3	3	2	1	2
206	B	3	1	1	5	1	1	3	1	1	3	4	3	2
210	A	1	1	1	5	4	3	3	2	3	3	2	4	2
210	B	1	2	2	2	4	2	4	2	2	3	4	1	2
221	A	1	5	2	2	3	2	1	2	2	3	3	1	3
221	B	3	4	2	5	2	2	1	1	4	4	3	3	3
263	A	5	1	2	2	3	2	3	2	3	2	2	1	1
263	B	4	2	3	4	2	2	3	2	1	4	2	3	1
264	A	3	5	1	2	2	4	3	4	3	3	2	2	1
264	B	1	5	2	2	4	4	2	3	1	3	2	3	2
268	A	2	4	3	5	3	4	3	4	3	2	1	3	2
268	B	2	1	4	5	3	3	3	1	2	1	1	2	1
272	A	2	4	4	4	4	2	3	4	4	4	2	3	2
272	B	2	4	3	4	4	4	2	4	1	3	1	1	1

8.12. Appendix 12: EQ-5D-5L and ASCOT valuation - Survey content

Page 1 - Initial screening demographics

Please indicate your age group

- 18 – 29 ☐
- 30 – 44 ☐
- 45 – 59 ☐
- 60 – 74 ☐
- 75+ ☐

What is your gender?

- Male ☐
- Female ☐

Page 2 - Survey information

Title: Welcome to the health comparison study

We are inviting you to participate in a study designed to gain an understanding of people's opinions about health and social care related quality of life. This will help decision makers in Australia and internationally to focus on the areas that Australian's value the highest. Your responses to hypothetical scenarios will be used to help us understand what is most important to people in making decisions about different health care treatments that aim to improve quality of life or length of life or both. This study is being undertaken by the Centre for Health Economics Research and Evaluation at the University of Technology Sydney in collaboration with Curtin University Perth. It has been ethically approved by the University of Technology Sydney.

Your participation in this study is completely voluntary. You are not obliged to participate and may stop at any time. Your responses to the survey are strictly confidential and at no time will the answers you give be linked to your identity. To complete this survey, we would expect someone to take approximately 15 minutes.

Page 3 – Survey information cont'd

Title – Survey information

This survey contains four sections

Section A briefly introduces the method we will use to describe health, and asks you to rate your own health.

Section B contains 15 questions. In each question you will be shown a number of possible health scenarios. You will then be asked to choose which you would prefer to experience. These profiles do not represent particular conditions and have been made up for the purpose of this exercise. Each option is described in terms of how good your health will be in a number of different areas.

Section C contains questions about you which will allow us to apply the results of the study to the population as a whole.

Section D is a brief feedback form about the questions you have completed.

Pages 4 – 16: Self-complete health-related and social related quality of life questions (EQ-5D-5L)

In this survey, we will describe health in a particular way, using a number of different areas such as mobility, pain/discomfort and anxiety/depression. To familiarise yourself with this approach, please answer these questions.

"Under each heading, please tick the ONE box that best describes your health TODAY"

Mobility

You have no problems in walking about
You have slight problems in walking about
You have moderate problems in walking about
You have severe problems in walking about
You are unable to walk about

Self-care

You have no problems washing and dressing yourself
You have slight problems washing and dressing yourself
You have moderate problems washing and dressing yourself
You have severe problems washing and dressing yourself
You are unable to wash and dress yourself

Usual activities

You have no problems doing your usual activities
You have slight problems doing your usual activities
You have moderate problems doing your usual activities
You have severe problems doing your usual activities
You are unable to do your usual activities

Pain/discomfort

You have no pain or discomfort
You have slight pain or discomfort
You have moderate pain or discomfort
You have severe pain or discomfort
You have extreme pain or discomfort

Anxiety/depression

You are not anxious or depressed
You are slightly anxious or depressed
You are moderately anxious or depressed
You are severely anxious or depressed
You are extremely anxious or depressed

Self-complete health-related and social related quality of life questions (ASCOT)
In this survey, we will describe social care related quality of life in a particular way, using a number of different areas such as control, safety and independence. To familiarise yourself with this approach, please answer these questions.

Which of the following statements best describes how much control you have over your daily life?
By 'control over daily life' we mean having the choice to do things or have things done as you like and when you want
You have as much control over your daily life as you want
You have adequate control over your daily life
You have some control over your daily life, but not enough
You have no control over your daily life

Thinking about keeping clean and presentable in appearance, which of the following statements best describes your situation?
You feel clean and are able to present yourself the way you like
You feel adequately clean and presentable
You feel less than adequately clean or presentable
You don't feel at all clean or presentable

Thinking about the food and drink you get, which of the following statements best describes your situation?
You get all the food and drink you like when you want
You get adequate food and drink at okay times
You don't always get adequate or timely food and drink
You don't always get adequate or timely food and drink, and think there is a risk to your health

Which of the following statements best describes how safe you feel?
By 'feeling safe' we mean how safe you feel both inside and outside the home. This includes fear of abuse, falling or other physical harm
You feel as safe as you want
You feel adequately safe, but not as safe as you would like
You feel less than adequately safe
You don't at all feel safe

Thinking about how much contact you have with people you like, which of the following statements best describes your social situation?
You have as much social contact as you want with people you like
You have adequate social contact with people
You have some social contact with people, but not enough
You have little social contact with people and feel socially isolated

Which of the following statements best describes how you spend your time?
You are able to spend time as you want, doing things you value or enjoy

You are able to do enough of the things you value or enjoy with your time
You do some of the things you value or enjoy with your time, but not enough
You don't do anything you value or enjoy with your time

Which of the following statements best describes how clean and comfortable your home is?
Your home is as clean and comfortable as you want
Your home is adequately clean and comfortable
Your home is not quite clean or comfortable enough
Your home is not at all clean or comfortable

Which of these statements best describes how the way you are helped and treated makes you think and feel about yourself?
The way you are helped and treated makes you think and feel better about yourself
The way you are helped and treated does not affect the way you think or feel about yourself
The way you are helped and treated sometimes undermines the way you think and feel about yourself
The way you are helped and treated completely undermines the way you think and feel about yourself

Page 17: Instruction page

Title: Section B – Making choices between options

You will now be presented with 15 questions.

In this set of questions, you will see two different descriptions of health and social care. Your task is to imagine living with the problems described. Then tell us which of the descriptions you would prefer to live in.

Please remember that there are no right or wrong answers. Some of the health descriptions may be difficult for you to imagine - just do the best you can. We are interested in your views, because it will help us to understand what aspects of quality of life are most important to people.

Let's start the questions now.

Pages 18 – 32: Example DCE question (Dimension order is EQ-5D-5L – ASCOT)

Please consider and imagine living with the two health descriptions below. Then tell us which description you would prefer to live in.

Description A	Description B
You have moderate problems in walking about	You are unable to walk about
You have severe problems washing and dressing yourself	You have slight problems washing and dressing yourself
You have moderate problems doing your usual activities	You have severe problems doing your usual activities
You have extreme pain or discomfort	You have slight pain or discomfort
You are severely anxious or depressed	You are moderately anxious or depressed
You have adequate control over your daily life	You have some control over your daily life, but not enough
You feel clean and are able to present yourself the way you like	You feel clean and are able to present yourself the way you like
You get adequate food and drink at okay times	You don't always get adequate or timely food and drink, and think there is a risk to your health
You feel adequately safe, but not as safe as you would like	You feel adequately safe, but not as safe as you would like
You have adequate social contact with people	You have some social contact with people, but not enough
You don't do anything you value or enjoy with your time	You don't do anything you value or enjoy with your time
Your home is adequately clean and comfortable	Your home is adequately clean and comfortable
The way you are helped and treated makes you think and feel better about yourself	The way you are helped and treated makes you think and feel better about yourself
Which do you prefer?	
Health Description A	**Health Description B**

Pages 18 - 32: Example DCE question (Dimension order is ASCOT – EQ-5D-5L)

Health description A	Health description B
You have adequate control over your daily life	You have some control over your daily life, but not enough
You feel clean and are able to present yourself the way you like	You feel clean and are able to present yourself the way you like
You get adequate food and drink at okay times	You don't always get adequate or timely food and drink, and think there is a risk to your health
You feel adequately safe, but not as safe as you would like	You feel adequately safe, but not as safe as you would like
You have adequate social contact with people	You have some social contact with people, but not enough
You don't do anything you value or enjoy with your time	You don't do anything you value or enjoy with your time
Your home is adequately clean and comfortable	Your home is adequately clean and comfortable
The way you are helped and treated makes you think and feel better about yourself	The way you are helped and treated makes you think and feel better about yourself
You have moderate problems in walking about	You are unable to walk about
You have severe problems washing and dressing yourself	You have slight problems washing and dressing yourself
You have moderate problems doing your usual activities	You have severe problems doing your usual activities
You have extreme pain or discomfort	You have slight pain or discomfort
You are severely anxious or depressed	You are moderately anxious or depressed
Which do you prefer?	
Health description A	**Health description B**

Page 33: Further demographic questions (1)

You have now finished the health description questions. Thank you.
We now need to collect some data about you. This is to ensure we have a good spread of respondents.

What is your country of birth?

- Australia ☐
- United Kingdom ☐
- New Zealand ☐
- Italy ☐
- Vietnam ☐
- Greece ☐
- Germany ☐
- Philippines ☐
- China ☐
- Indonesia ☐
- Other ☐

If other, please specify_____

What is the highest level of education you completed?

- Primary ☐
- Secondary ☐
- Trade certificate/Diploma ☐
- Bachelor's degree ☐
- Higher degree ☐

Are you currently studying?

- Yes ☐
- No ☐

Page 34: Further demographic questions (2)

Please indicate which of these categories best matches your gross (before tax) income?

- Under $20,000 ☐
- $20,001-$30,000 ☐
- $30,001-$40,000 ☐
- $40,001-$50,000 ☐
- $50,001-$60,000 ☐
- $60,001-$70,000 ☐
- $70,001-$80,000 ☐
- $80,001-$100,000 ☐
- Over $100,000 ☐
- Prefer not to say ☐

What is your current marital status?

- Single ☐
- Separated/divorced ☐
- Widowed ☐
- Married/De facto ☐

How many children or dependents do you have?

- None ☐
- 1 ☐
- 2 ☐
- 3 or more ☐

Do you have experience of serious illness in:

	Yes	No
Yourself	☐	☐
Your family	☐	☐
Caring for others	☐	☐

Page 35: Self-report health questions

In general, would you say your health is?

Excellent ☐

Very good ☐

Good ☐

Fair ☐

Poor ☐

Do you have any illness, health problem, condition or disability?

Yes ☐

No ☐

If yes please tick all that apply:

Tiredness/fatigue		High blood pressure	
Pain		Heart disease	
Insomnia		Osteoarthritis	
Anxiety/nerves		Stroke	
Depression		Cancer	
Diabetes		Other	
Breathing problems (e.g. asthma, emphysema)			

Have you had a health condition requiring hospitalisation in the last five years?

Yes ☐

No ☐

Page 36: General feedback questions and free-text

Please indicate your level of agreement with the following statements about the tasks in general

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I found the tasks difficult					
I found it difficult to tell the difference between the descriptions					
I found it difficult to imagine the scenarios					
I considered the whole description whilst completing the task					

If you want to provide any further comments about the questions or survey in general, or the answers you provided please do so below.

FREE-TEXT HERE

Page 37: Final page

Thank you for taking part in the health comparison study

Please click the link to complete the survey

8.13. Appendix 13: HRQoL and SCRQoL interaction models

Table 75: Further exploratory interactions

Parameter	Model 51: All interactions of high HRQoL and SCRQoL and low dimension levels	
	Coef. (p)	SE
Main Effects		
MO2	0.055	0.146
MO3	-0.068	0.142
MO4	-0.418**	0.143
MO5	-0.642***	0.143
SC2	-0.117*	0.124
SC3	-0.343**	0.124
SC4	-0.613***	0.128
SC5	-0.634***	0.127
UA2	-0.025	0.141
UA3	-0.007	0.142
UA4	-0.265**	0.144
UA5	-0.270**	0.143
PD2	-0.304**	0.120
PD3	-0.304**	0.120
PD4	-0.717***	0.120
PD5	-0.871***	0.121
AD2	-0.103	0.143
AD3	-0.261	0.144
AD4	-0.625***	0.139
AD5	-0.763***	0.143
CO2	-0.523***	0.136
CO3	-0.621***	0.137
CO4	-1.018***	0.137
CL2	0.178	0.132
CL3	0.104	0.132
CL4	0.026	0.132
FD2	-0.101	0.142
FD3	-0.242*	0.142
FD4	-0.316**	0.141
SA2	-0.075	0.131
SA3	-0.145	0.127
SA4	-0.339**	0.127
SP2	0.243	0.150
SP3	0.223	0.152
SP4	0.118	0.147
OC2	-0.169	0.134
OC3	-0.183	0.131
OC4	-0.323**	0.132
AC2	0.109	0.176
AC3	0.015	0.175
AC4	-0.093	0.174
DI2	-0.171	0.133
DI3	-0.220	0.133
DI4	-0.289*	0.133
Interactions		
ASCOT Level 4	-0.140	0.098
MO1 x ASCOT Level 4	0.183	0.145
SC1 x ASCOT Level 4	-0.123*	0.129
UA1 x ASCOT Level 4	-0.008	0.146
PD1 x ASCOT Level 4	-0.039	0.123
AD1 x ASCOT Level 4	-0.059	0.144
EQ-5D Levels 4/5	-0.054	0.101
CO1 x EQ-5D Levels 4/5	-0.388**	0.138
CL1 x EQ-5D Levels 4/5	0.326**	0.130

FD1 x EQ-5D Levels 4/5	0.012	0.145
SA1 x EQ-5D Levels 4/5	-0.008	0.128
SP1 x EQ-5D Levels 4/5	0.271	0.153
OC1 x EQ-5D Levels 4/5	-0.090	0.135
AC1 x EQ-5D Levels 4/5	0.095	0.173
DI1 x EQ-5D Levels 4/5	-0.171	0.134
No Obs	14,625	
AIC	17,985	
BIC	18,473	

8.14. Appendix 14: Scale testing for demographic variables

Table 76 to **Table 78** report the results of the scale testing analysis for gender, age category, and condition status respectively

Table 76: Conditional logit and pooled models by gender

Parameter	Model 52: Male		Model 53: Female		Model 54: Restricted pooled	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.172**	0.065	-0.061	0.066	-0.073*	0.033
MO3	-0.236***	0.065	-0.267***	0.066	-0.170***	0.035
MO4	-0.651***	0.064	-0.553***	0.063	-0.408***	0.045
MO5	-0.847***	0.070	-0.758***	0.070	-0.546***	0.054
SC2	**0.018**	0.062	-0.020	0.062	-0.001	0.030
SC3	-0.258***	0.063	-0.138*	0.063	-0.132***	0.033
SC4	-0.455***	0.061	-0.511***	0.064	-0.333***	0.039
SC5	**0.450*** **	0.064	-0.588***	0.066	-0.360***	0.041
UA2	-0.070	0.067	**0.025**	0.070	-0.011	0.033
UA3	**0.003**	0.066	-0.052	0.068	-0.019	0.033
UA4	-0.194**	0.067	-0.362***	0.070	-0.198***	0.036
UA5	-0.212***	0.066	**-0.341*** **	0.067	**-0.191*** **	0.035
PD2	-0.242****	0.067	-0.317***	0.070	-0.193***	0.036
PD3	**-0.155***	0.069	-0.361***	0.072	**-0.184*** **	0.037
PD4	-0.546***	0.067	-0.854***	0.071	-0.492***	0.048
PD5	-0.741***	0.064	-0.962***	0.067	-0.591***	0.054
AD2	-0.052	0.065	-0.033	0.065	-0.024	0.032
AD3	-0.126	0.068	-0.276***	0.070	-0.141***	0.035
AD4	-0.475***	0.068	-0.678***	0.067	-0.401***	0.044
AD5	-0.610***	0.067	-0.819***	0.069	-0.496***	0.049
CO2	-0.139	0.059	-0.188**	0.060	-0.114***	0.030
CO3	-0.184**	0.060	-0.311***	0.061	-0.175***	0.032
CO4	-0.581***	0.059	-0.769***	0.059	-0.468***	0.044
CL2	-0.054	0.062	-0.169**	0.063	-0.082**	0.031
CL3	-0.086	0.060	-0.295***	0.062	-0.138***	0.031
CL4	-0.184**	0.062	-0.402***	0.064	-0.211***	0.034
FD2	-0.064	0.063	-0.135*	0.064	-0.072*	0.031
FD3	-0.210***	0.061	-0.306***	0.062	-0.182***	0.033
FD4	-0.316***	0.066	-0.403***	0.066	-0.251***	0.037
SA2	-0.037	0.062	-0.081	0.062	-0.042	0.030
SA3	-0.083	0.059	-0.165**	0.061	-0.090**	0.034
SA4	-0.249***	0.059	-0.412***	0.060	-0.234***	0.034
SP2	**0.083**	0.060	-0.105	0.062	-0.015	0.030
SP3	-0.010	0.064	**-0.091**	0.065	-0.035	0.031
SP4	-0.053	0.059	-0.243***	0.059	-0.108***	0.030
OC2	-0.107	0.065	-0.057	0.057	-0.054	0.032
OC3	**-0.096**	0.061	**-0.050**	0.063	**-0.048**	0.031
OC4	-0.212	0.060	-0.260***	0.061	-0.162***	0.032
AC2	**0.042**	0.060	**0.022**	0.059	**0.023**	0.030
AC3	-0.083	0.059	-0.040	0.063	-0.037	0.029
AC4	-0.207	0.061	-0.204***	0.063	-0.138***	0.032
DI2	**0.046**	0.056	-0.063	0.059	-0.008	0.028
DI3	**0.013**	0.061	-0.151*	0.064	-0.051	0.031
DI4	-0.071	0.059	-0.257***	0.062	-0.116***	0.030
Gender					0.237***	0.047
No obs[c]	7,215		7,410		14,625	
LL[d]	-4,512		-4,398		-8,936	
AIC[e]	9,113		8,884		17,962	
BIC[f]	9,446		9,219		18,334	

[a] Coefficient estimate; b standard error; c number of observations; d Log-Likelihood; e Akaike Information Criterion; f Bayesian Information Criterion; p-values for difference between coefficient estimate and baseline indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at 0.05; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

Table 77: Conditional logit and pooled models by age

Parameter	Model 55: Under 65		Model 56: 65 or over		Model 57: Restricted pooled	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.150**	0.053	**0.006**	0.098	-0.086*	0.042
MO3	-0.243***	0.053	-0.281**	0.098	-0.221***	0.042
MO4	-0.573***	0.051	-0.717***	0.096	-0.538***	0.042
MO5	-0.723***	0.056	-1.115***	0.107	-0.736***	0.046
SC2	-0.016	0.050	**0.051**	0.092	0.001	0.039
SC3	-0.181***	0.051	-0.274**	0.094	-0.181***	0.040
SC4	-0.416***	0.050	-0.741***	0.093	-0.450***	0.040
SC5	-0.446***	0.052	-0.800***	0.097	-0.482***	0.042
UA2	-0.053	0.055	**0.127**	0.103	-0.004	0.043
UA3	-0.068	0.054	**0.139**	0.100	-0.011	0.043
UA4	-0.287***	0.055	-0.222*	0.102	-0.242***	0.044
UA5	**-0.269*** **	0.053	-0.264**	0.102	-0.245***	0.043
PD2	-0.322***	0.055	-0.097	0.105	-0.230***	0.044
PD3	**-0.291*** **	0.057	-0.141	0.107	-0.218***	0.045
PD4	-0.717***	0.055	-0.632***	0.104	-0.615***	0.046
PD5	-0.906***	0.053	-0.865***	0.096	-0.742***	0.045
AD2	-0.057	0.052	**0.015**	0.100	-0.036	0.041
AD3	-0.233***	0.055	-0.099	0.103	-0.170***	0.044
AD4	-0.570***	0.054	-0.629***	0.102	-0.519***	0.044
AD5	-0.672***	0.054	-0.865***	0.104	-0.647***	0.045
CO2	-0.130**	0.048	-0.270**	0.090	-0.147	0.038
CO3	-0.233***	0.048	-0.277**	0.090	-0.218	0.039
CO4	-0.565***	0.047	-1.015***	0.090	-0.615	0.039
CL2	-0.120**	0.050	-0.090	0.095	-0.098	0.040
CL3	-0.126**	0.049	-0.419***	0.091	-0.186	0.039
CL4	-0.265***	0.051	-0.422***	0.095	-0.273	0.040
FD2	-0.103*	0.051	-0.103	0.096	-0.095	0.040
FD3	-0.221***	0.049	-0.413***	0.093	-0.245	0.039
FD4	-0.343***	0.053	-0.477***	0.099	-0.331	0.042
SA2	-0.019	0.050	-0.169	0.092	-0.058	0.039
SA3	-0.089	0.048	-0.247**	0.089	-0.125	0.038
SA4	-0.311***	0.048	-0.426***	0.090	-0.302	0.038
SP2	-0.020	0.049	**0.041**	0.090	-0.004	0.039
SP3	-0.047	0.052	-0.026	0.095	-0.039	0.041
SP4	-0.091	0.048	-0.315***	0.088	-0.140***	0.037
OC2	-0.089	0.052	-0.061	0.099	-0.068	0.042
OC3	-0.051	0.050	-0.162	0.093	-0.070	0.039
OC4	-0.200***	0.049	-0.380***	0.092	-0.222***	0.039
AC2	0.051	0.049	**0.015**	0.092	0.029	0.039
AC3	-0.040	0.047	-0.077	0.087	-0.054	0.037
AC4	-0.139***	0.050	-0.388***	0.092	-0.196***	0.039
DI2	-0.045	0.046	**0.130**	0.087	0.009	0.037
DI3	-0.062	0.050	-0.034	0.094	-0.052	0.040
DI4	-0.148***	0.049	-0.174*	0.088	-0.142***	0.038
Age					0.352***	0.050
No obs[c]	10,995		3,630		14,625	
LL[d]	-6,823		-2,056		-8,926	
AIC[e]	13,734		4,201		17,943	
BIC[f]	14,086		4,505		18,316	

[a] Coefficient estimate; b standard error; c number of observations; d Log-Likelihood; e Akaike Information Criterion; f Bayesian Information Criterion; p-values for difference between coefficient estimate and baseline indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at 0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

Table 78: Conditional logit and pooled models by condition status

Parameter	Model 58: No condition		Model 59: Has condition		Model 60: Restricted pooled	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.156**	0.061	-0.054	0.071	-0.096*	0.042
MO3	-0.272***	0.062	-0.207**	0.071	-0.222***	0.043
MO4	-0.629***	0.059	-0.560***	0.069	-0.541***	0.044
MO5	-0.758***	0.065	-0.863***	0.077	-0.731***	0.048
SC2	-0.023	0.058	0.023	0.067	**0.001**	0.040
SC3	-0.208***	0.059	-0.189**	0.068	-0.177***	0.041
SC4	-0.469***	0.058	-0.498***	0.068	-0.439***	0.042
SC5	-0.481***	0.061	-0.562***	0.070	-0.473***	0.043
UA2	-0.055	0.063	**0.034**	0.074	-0.014	0.044
UA3	**-0.023**	0.062	-0.031	0.072	-0.022	0.043
UA4	-0.267***	0.064	-0.294***	0.073	-0.254***	0.044
UA5	**-0.239*** **	0.062	-0.315***	0.072	**-0.252*** **	0.043
PD2	-0.296***	0.064	-0.241***	0.074	-0.247***	0.045
PD3	-0.343***	0.066	**-0.144**	0.076	**-0.225*** **	0.046
PD4	-0.654***	0.064	-0.751***	0.074	-0.636***	0.047
PD5	-0.882***	0.061	-0.820***	0.070	-0.770***	0.046
AD2	-0.036	0.060	-0.039	0.071	-0.036	0.042
AD3	-0.182**	0.064	-0.216**	0.074	-0.183***	0.045
AD4	-0.537***	0.063	-0.626***	0.073	-0.528***	0.045
AD5	-0.644***	0.064	-0.788***	0.074	-0.655***	0.046
CO2	-0.119*	0.055	-0.213***	0.065	0.148***	0.039
CO3	-0.260***	0.056	-0.223***	0.065	-0.222***	0.039
CO4	-0.560***	0.055	-0.822***	0.064	-0.616***	0.040
CL2	-0.081	0.058	-0.165*	0.068	-0.103**	0.040
CL3	-0.117*	0.057	-0.293***	0.066	-0.179***	0.039
CL4	-0.250***	0.059	-0.367***	0.068	-0.273***	0.041
FD2	-0.120*	0.060	-0.086	0.068	-0.093*	0.041
FD3	-0.206***	0.057	-0.342***	0.066	-0.243***	0.040
FD4	-0.374***	0.062	-0.356***	0.071	-0.330***	0.043
SA2	-0.096	0.058	**0.003**	0.066	-0.047	0.040
SA3	-0.167**	0.056	-0.074	0.064	-0.112**	0.039
SA4	-0.381***	0.056	-0.262***	0.064	-0.296***	0.039
SP2	-0.027	0.057	**0.012**	0.066	-0.007	0.039
SP3	**-0.014**	0.060	-0.095	0.070	-0.045	0.042
SP4	-0.090	0.055	-0.221***	0.064	-0.138***	0.038
OC2	-0.059	0.061	-0.119	0.071	-0.077	0.042
OC3	**-0.021**	0.058	-0.138*	0.067	**-0.071**	0.040
OC4	-0.186***	0.057	-0.310***	0.066	-0.221***	0.039
AC2	0.081	0.057	-0.034	0.066	**0.026**	0.039
AC3	**-0.037**	0.055	-0.093	0.063	-0.053	0.038
AC4	-0.120*	0.058	-0.311***	0.067	-0.191***	0.040
DI2	-0.016	0.054	**0.012**	0.063	-0.004	0.037
DI3	-0.021	0.058	-0.114	0.068	-0.061	0.040
DI4	-0.114*	0.056	-0.222***	0.066	-0.150***	0.039
Condition status					0.188***	0.046
No obs[c]	8,160		6,465		14,625	
LL[d]	-5,062		-3,852		-8,941	
AIC[e]	10,212		7,792		17,972	
BIC[f]	10,551		8,121		18,344	

[a] Coefficient estimate; b standard error; c number of observations; d Log-Likelihood; e Akaike Information Criterion; f Bayesian Information Criterion; p-values for difference between coefficient estimate and baseline indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at 0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

8.15. Appendix 15: Time taken sensitivity analysis

The task subset exclusion models are displayed in **Figure 39**. Task subset 1 excludes 602 (4.1%) task completions, with 591 (4.0%) completed in less than 4 seconds, and 11 (<0.01%) completed in more than 30 minutes. Therefore, the subset includes 14,023 tasks and has the lowest number of inconsistent coefficient estimates (3). This demonstrates that removing the outliers at each end of the completion time scale may increase model validity. Task subset 2 excludes 1,312 (9.0%) of the tasks with 591 (4.0%) completed in less than four seconds, and 721 (4.9%) completed in more than 92 seconds. This model is based on 13,313 task completions, and results in five inconsistencies. Subset 3 removes 7,344 (50.2%) of the tasks with 3,808 (26.0%) completed in less than 9 seconds, and 3,536 (24.2%) completed in more than 37 seconds. This model results in four inconsistencies. Subset four is based on the 7,334 observations excluded for set 3. This model results in eight inconsistencies which suggests that tasks completed in a shorter and longer time overall contribute to increased disordering across coefficient levels.

Figure 39: Sensitivity analysis of time taken per task

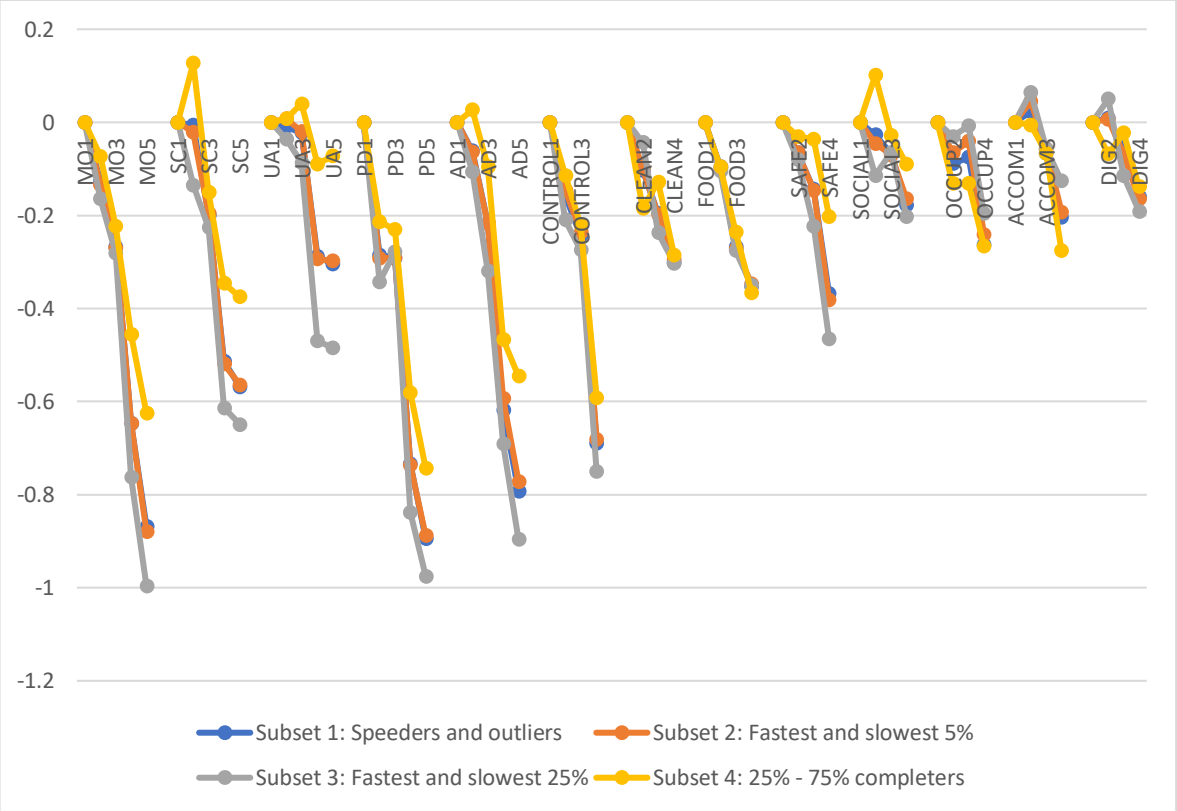
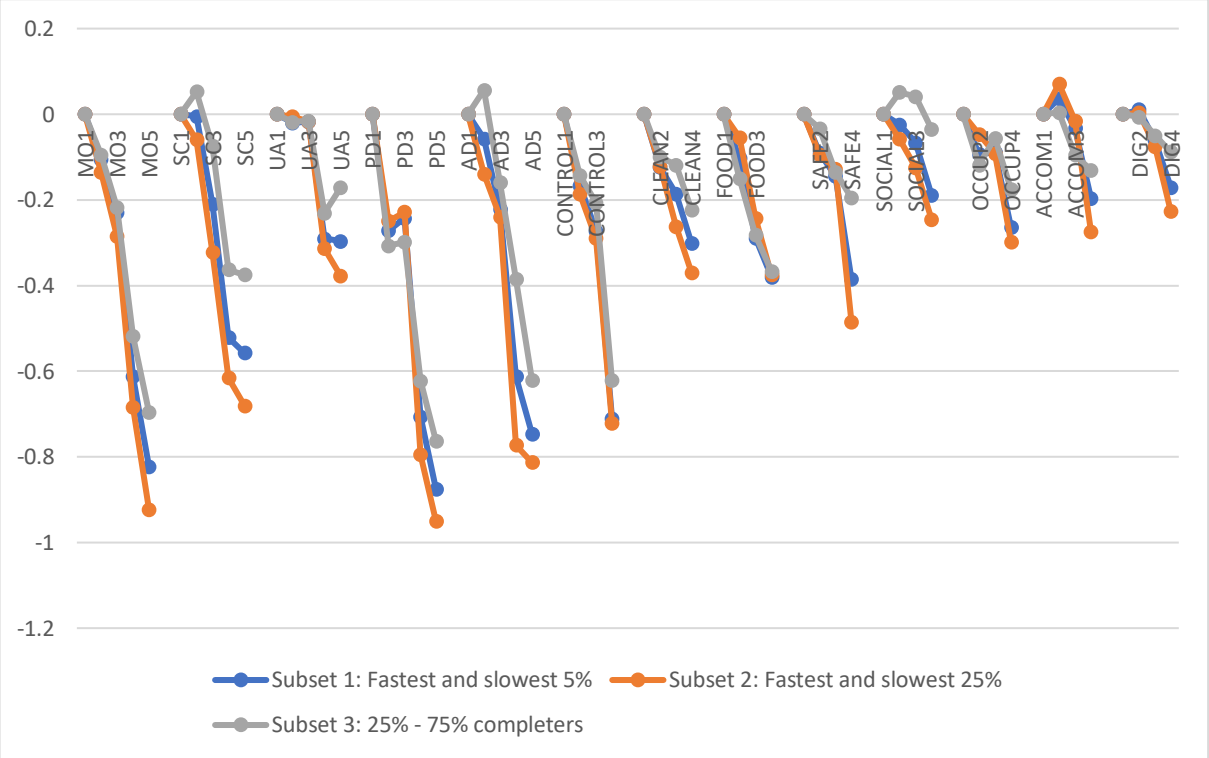


Figure 40 displays the impact on the estimates of removing respondents based on the overall time taken to complete the survey. Respondent subset 1 excludes 98 (10.1%) respondents completing the survey in below 8 minutes (48; 4.9%), or above 66 minutes (50; 5.1%). Therefore,

the model includes 877 (90.0%) respondents and results in five inconsistencies. Subset 2 removes 477 (48.9%) respondents completing the survey in below 13 minutes (242; 24.8%) or above 32 minutes (235; 21.1%). The model therefore includes 498 (51.1%) respondents completing the survey in between 25% and 75% of the overall time taken, and results in three inconsistencies. Subset 3 includes the 477 (48.9%) respondents removed from set two, and results in nine inconsistencies. This is a similar pattern to the exclusions based on task time in that respondents completing the survey in a shorter or longer time overall contribute to increased disordering across coefficient levels.

Figure 40: Sensitivity analysis of time taken to complete survey



Differences in preference patterns based on the time taken to complete the tasks and the survey were also assessed using scale testing. **Table 79** reports the results of the conditional logit model for two subsamples divided by the median time taken to complete the tasks, as well as the pooled scale model. The model including tasks completed in less than the median time (7,344 (50.2%) observations) resulted in 12 inconsistencies, in comparison to the model including the 7,281 (49.2%) of tasks completed in more than the median time which resulted in five inconsistencies. The LR statistic is 96.6, which is higher than the critical value (61.7) meaning that the null hypothesis of preference homogeneity is rejected, and scale differs according to the time taken to complete the tasks.

Table 80 includes the models based on dividing the sample based on the overall median completion time. The model including the 488 (50.1%) completing the survey in less than the median time resulted in seven inconsistencies, in comparison to the model including the 487 (49.9%) above the median which resulted in three. In contrast to the task time model, the LR statistic was 58.2, so below the critical value meaning that the null hypothesis of preference homogeneity is accepted.

Table 79: Conditional logit and pooled models by median time per task

Parameter	Model 61: Less than median		Model 62: Above median		Model 63: Restricted pooled	
	Coef. (p)	SE	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.120	0.064	-0.075	0.069	-0.027	0.015
MO3	-0.310***	0.064	-0.185**	0.070	-0.070***	0.016
MO4	-0.563***	0.062	-0.663***	0.067	-0.184***	0.022
MO5	-0.724***	0.067	-0.944***	0.076	-0.261***	0.028
SC2	-0.027	0.061	-0.020	0.065	**-0.005**	0.014
SC3	-0.152*	0.061	-0.309***	0.067	-0.071***	0.015
SC4	-0.382***	0.060	-0.678***	0.067	-0.168***	0.020
SC5	**-0.317*** **	0.062	-0.806***	0.070	-0.184***	0.021
UA2	**0.056**	0.066	-0.105	0.073	-0.012	0.015
UA3	**0.028**	0.064	**-0.079**	0.072	**-0.011**	0.015
UA4	-0.127	0.067	-0.441***	0.072	-0.099***	0.018
UA5	**-0.083**	0.064	-0.517***	0.071	-0.108***	0.017
PD2	-0.214***	0.065	-0.312***	0.075	-0.085***	0.017
PD3	**-0.276*** **	0.068	**-0.262*** **	0.075	**-0.080*** **	0.018
PD4	-0.491***	0.065	-0.957***	0.075	-0.235***	0.026
PD5	-0.718***	0.063	-1.035***	0.069	-0.271***	0.028
AD2	-0.025	0.062	-0.060	0.070	-0.014	0.015
AD3	-0.186**	0.066	-0.289***	0.074	-0.071***	0.017
AD4	-0.368***	0.065	-0.812***	0.072	-0.194***	0.023
AD5	-0.537***	0.065	-0.973***	0.074	-0.241***	0.026
CO2	-0.070	0.057	-0.271***	0.065	-0.057***	0.014
CO3	-0.125*	0.058	-0.374***	0.065	-0.082***	0.015
CO4	-0.430***	0.055	-0.960***	0.064	-0.226***	0.024
CL2	-0.052	0.061	-0.173**	0.066	-0.040**	0.014
CL3	-0.025	0.060	-0.389***	0.063	-0.077***	0.015
CL4	-0.071	0.062	-0.542***	0.066	-0.113***	0.017
FD2	-0.141*	0.061	-0.027	0.067	-0.021	0.015
FD3	-0.191***	0.060	-0.324***	0.064	-0.082***	0.016
FD4	-0.272***	0.066	-0.418***	0.068	-0.109***	0.018
SA2	**0.005**	0.059	**-0.139***	0.067	-0.025	0.014
SA3	-0.044	0.058	-0.232***	0.064	-0.047***	0.014
SA4	-0.126*	0.057	-0.579***	0.065	-0.123***	0.017
SP2	-0.077	0.059	**0.084**	0.065	**0.007**	0.014
SP3	**-0.042**	0.061	-0.038	0.070	-0.013	0.015
SP4	**-0.010**	0.056	-0.273***	0.062	-0.054***	0.014
OC2	-0.017	0.063	-0.114	0.069	-0.024	0.015
OC3	**0.037**	0.061	-0.191**	0.065	-0.032*	0.014
OC4	-0.054	0.058	-0.453***	0.066	-0.091***	0.015
AC2	**0.045**	0.059	**0.002**	0.064	**0.008**	0.014
AC3	**-0.000**	0.057	-0.165**	0.061	-0.029*	0.013
AC4	-0.090	0.060	-0.370***	0.067	-0.077***	0.015
DI2	**0.002**	0.055	-0.006	0.062	-0.000	0.013
DI3	-0.009	0.060	-0.116	0.067	-0.021	0.014
DI4	-0.035	0.059	-0.319***	0.064	-0.064***	0.014
Time					0.709***	0.051
No obs	7,344		7,281		14,625	
LL	-4,731		-4,057		-8,836	
AIC	9,550		8,202		17,763	
BIC	9,884		8,536		18,136	

[a] Coefficient estimate; b standard error; c number of observations; d Log-Likelihood; e Akaike

Information Criterion; f Bayesian Information Criterion; p-values for difference between coefficient estimate and baseline indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at 0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

Table 80: Conditional logit and pooled models by median overall completion time

Parameter	Model 64: Less than median		Model 65: Above median		Model 66: Restricted pooled	
	Coef. (p)[a]	SE[b]	Coef. (p)	SE	Coef. (p)	SE
MO2	-0.085	0.063	-0.130	0.068	-0.086*	0.036
MO3	-0.260***	0.064	-0.223***	0.068	-0.184***	0.037
MO4	-0.505***	0.062	-0.701***	0.066	-0.473***	0.038
MO5	-0.722***	0.068	-0.882***	0.072	-0.623***	0.044
SC2	-0.093	0.060	0.101	0.064	**0.014**	0.034
SC3	-0.231***	0.062	-0.176**	0.065	-0.150***	0.035
SC4	-0.476***	0.060	-0.503***	0.065	-0.374***	0.037
SC5	**-0.400*** **	0.062	-0.661***	0.068	-0.419***	0.038
UA2	**0.051**	0.066	-0.099	0.071	-0.025	0.037
UA3	-0.025	0.065	**-0.017**	0.069	-0.016	0.037
UA4	-0.194	0.066	-0.369	0.071	-0.225***	0.039
UA5	**-0.128**	0.065	-0.429	0.069	-0.232***	0.037
PD2	-0.264***	0.066	-0.282***	0.072	-0.213***	0.039
PD3	**-0.219*** **	0.069	-0.301***	0.073	**-0.206*** **	0.040
PD4	-0.579***	0.066	-0.836***	0.072	-0.556***	0.042
PD5	-0.746***	0.063	-0.960***	0.068	-0.665***	0.042
AD2	-0.038	0.063	-0.041	0.067	-0.029	0.036
AD3	-0.110	0.066	-0.302***	0.072	-0.164***	0.038
AD4	-0.382***	0.065	-0.789***	0.070	-0.468***	0.040
AD5	-0.525***	0.066	-0.919***	0.071	-0.571***	0.041
CO2	-0.191***	0.058	-0.141*	0.061	-0.122***	0.033
CO3	-0.233***	0.058	-0.275***	0.063	-0.193***	0.034
CO4	-0.583***	0.057	-0.762***	0.061	-0.522***	0.037
CL2	-0.061	0.061	-0.162*	0.065	-0.093**	0.035
CL3	-0.114*	0.059	-0.277***	0.063	-0.158***	0.034
CL4	-0.215***	0.062	-0.385***	0.065	-0.239***	0.035
FD2	-0.133*	0.062	-0.061	0.065	-0.067	0.035
FD3	-0.263***	0.060	-0.258***	0.063	-0.197***	0.035
FD4	-0.359***	0.064	-0.366***	0.068	-0.276***	0.038
SA2	-0.068	0.060	**-0.049**	0.064	-0.044	0.035
SA3	-0.140*	0.058	-0.115	0.062	-0.096**	0.033
SA4	-0.256***	0.058	-0.421***	0.062	-0.267***	0.034
SP2	-0.063	0.060	**0.047**	0.063	**0.002**	0.034
SP3	**-0.021**	0.062	-0.068	0.067	-0.036	0.036
SP4	-0.071	0.058	-0.224***	0.060	-0.122***	0.033
OC2	-0.090	0064	-0.071	0.067	-0.061	0.036
OC3	**-0.011**	0.061	-0.178***	0.064	**-0.074***	0.034
OC4	-0.158**	0.059	-0.323***	0.063	-0.195***	0.034
AC2	0.087	0.059	-0.020	0.063	**0.020**	0.034
AC3	**-0.047**	0.057	-0.059	0.060	-0.041	0.032
AC4	-0.123*	0.060	-0.291***	0.065	-0.165***	0.035
DI2	0.019	0.056	-0.020	0.060	-0.003	0.032
DI3	0.020	0.061	-0.150*	0.065	-0.059	0.035
DI4	-0.006	0.058	-0.318***	0.063	-0.142***	0.033
Time					0.428***	0.048
No obs[c]	7,320		7,305		14,625	
LL[d]	-4,654		-4,222		-8,906	
AIC[e]	9,397		8,533		17,903	
BIC[f]	9,731		8,867		18,276	

[a] Coefficient estimate; b standard error; c number of observations; d Log-Likelihood; e Akaike Information Criterion; f Bayesian Information Criterion; p-values for difference between coefficient estimate and baseline indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at

0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

8.16. Appendix 16: Latent class models with between three and six classes

Table 81: Three class latent class model (Model 67)

Parameter	Class 1	Class 2	Class 3
MO2	-0.096	-0.120	-0.110
MO3	-0.123	**-0.019**	-0.712
MO4	-0.497	-0.215	-1.281
MO5	-0.434	-0.302	-1.934
SC2	**0.231**	**0.005**	-0.126
SC3	-0.094	-0.121	-0.568
SC4	-0.235	-0.134	-1.367
SC5	-0.544	**0.104**	-1.430
UA2	-0.241	-0.031	**0.063**
UA3	**-0.092**	**0.051**	-0.054
UA4	-0.874	**0.030**	-0.349
UA5	**-0.784**	**0.083**	-0.450
PD2	-0.849	-0.196	-0.134
PD3	-0.907	**-0.148**	**-0.075**
PD4	-2.080	-0.152	-0.594
PD5	-2.417	-0.354	-0.550
AD2	-0.423	**0.000**	**0.219**
AD3	-0.756	-0.008	**-0.081**
AD4	-2.037	**0.137**	-0.491
AD5	-2.265	-0.093	-0.610
CO2	-0.208	-0.109	-0.310
CO3	-0.300	-0.218	**-0.288**
CO4	-0.975	-0.276	-1.247
CL2	-0.312	-0.076	-0.100
CL3	-0.301	**-0.045**	-0.413
CL4	-0.455	-0.134	-0.513
FD2	-0.075	-0.156	-0.064
FD3	-0.247	-0.196	-0.419
FD4	-0.415	-0.333	-0.448
SA2	-0.083	**0.097**	-0.278
SA3	-0.359	**0.124**	-0.372
SA4	-0.617	-0.007	-0.689
SP2	-0.061	**0.124**	-0.113
SP3	-0.137	**0.044**	-0.115
SP4	-0.464	**0.082**	-0.239
OC2	-0.223	**0.012**	-0.190
OC3	-0.219	**0.039**	**-0.180**
OC4	-0.202	-0.032	-0.646
AC2	**0.166**	-0.019	**0.096**
AC3	-0.231	-0.069	**0.104**
AC4	-0.254	-0.154	-0.231
DI2	-0.035	-0.014	**0.006**
DI3	-0.274	**0.027**	-0.026
DI4	-0.314	**0.036**	-0.287
Demographic			
Age Cat (18-60 and 60+)	-0.693	-1.477	0
Gender	0.107	-0.72	0
Has Long-term Condition	-0.034	-0.185	0
Class Share	0.309	0.330	0.361
N Obs[a]		14,625	
LL[b]		-8,607	

[a] Number of observations; [b] Log-Likelihood; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity;

Table 82: Four class latent class model (Model 68)

Parameter	Class 1	Class 2	Class 3	Class 4
MO2	-0.221	**0.297**	-0.768	-0.072
MO3	**-0.147**	-0.725	-1.20	**0.015**
MO4	-0.697	-0.669	-2.760	-0.133
MO5	-0.691	-1.532	-3.715	-0.181
SC2	**0.151**	**0.022**	-0.801	**0.035**
SC3	**0.032**	-0.589	-0.932	-0.104
SC4	-0.418	-1.690	-1.576	-0.160
SC5	-0.623	-1.955	**-1.508**	**0.078**
UA2	-0.130	**0.231**	-0.834	**0.041**
UA3	**-0.120**	**0.077**	**-0.564**	**0.077**
UA4	-0.871	-0.065	-1.534	**0.077**
UA5	**-0.763**	-0.169	-1.728	**0.144**
PD2	-0.918	**0.347**	-0.854	-0.169
PD3	-0.918	**0.184**	-0.972	**-0.078**
PD4	-2.093	-0.535	**-0.667**	**-0.131**
PD5	-2.406	**-0.406**	-1.157	-0.289
AD2	-0.445	**0.492**	**0.092**	**0.047**
AD3	-0.802	-0.201	-0.038	**0.064**
AD4	-1.857	-0.726	-0.336	**0.126**
AD5	-2.086	-1.203	**-0.007**	-0.106
CO2	-0.233	-0.284	-0.148	-0.121
CO3	-0.233	**-0.156**	-0.771	-0.188
CO4	-0.915	-1.950	**-0.696**	-0.333
CL2	-0.258	-0.156	-0.272	-0.084
CL3	**-0.244**	-0.565	-0.827	**0.003**
CL4	-0.482	**-0.524**	**-0.656**	-0.143
FD2	-0.028	-0.141	-0.400	-0.111
FD3	-0.201	-0.816	**-0.187**	-0.215
FD4	-0.348	**-0.605**	-1.018	-0.296
SA2	-0.001	-0.679	**0.059**	**0.060**
SA3	-0.259	-0.744	-0.041	**0.151**
SA4	-0.529	-1.225	-0.428	-0.002
SP2	-0.046	-0.295	**0.180**	**0.117**
SP3	-0.060	-0.346	**0.110**	**0.006**
SP4	-0.337	-0.696	**0.289**	**0.011**
OC2	-0.221	-0.171	-0.375	-0.008
OC3	**-0.197**	**-0.164**	-0.526	**0.018**
OC4	**-0.181**	-1.039	**-0.515**	-0.094
AC2	**0.143**	**0.045**	**0.181**	-0.024
AC3	-0.225	**0.017**	**0.455**	-0.086
AC4	**-0.119**	-0.585	**0.076**	-0.261
DI2	-0.062	-0.250	**0.140**	**0.015**
DI3	-0.262	-0.407	**0.338**	**0.019**
DI4	-0.318	-0.655	**0.118**	-0.001
Demographic				
Age Cat (18-60 and 60+)	0.578	1.867	1.205	0
Gender	0.727	0.994	0.045	0
Has Long-term Condition	0.086	0.131	-0.416	0
Class Share	0.341	0.194	0.131	0.333
N Obs[a]		14,625		
LL[b]		-8,862		

[a] Number of observations; [b] Log-Likelihood; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

Table 83: Five class latent class model (Model 69)

Parameter	Class 1	Class 2	Class 3	Class 4	Class 5
MO2	**0.229**	-1.318	-0.016	-0.791	**0.223**
MO3	**0.135**	**-0.991**	**0.073**	-1.056	-0.668
MO4	-0.187	-1.819	-0.045	-2.350	-0.790
MO5	-0.208	-1.580	-0.068	-3.256	-1.588
SC2	**0.420**	-0.403	**0.029**	-0.559	-0.029
SC3	**0.452**	-0.936	-0.074	-0.693	-0.744
SC4	**0.006**	-1.223	-0.173	-1.017	-1.820
SC5	-0.455	**-0.424**	**0.034**	-1.067	-2.051
UA2	-0.412	**0.644**	**0.033**	-0.675	**0.201**
UA3	-0.105	**0.004**	**0.055**	**-0.421**	**0.002**
UA4	-1.137	-0.308	**0.158**	-1.265	-0.110
UA5	-1.008	-0.121	**0.189**	-1.426	-0.207
PD2	-0.732	-0.955	-0.134	-0.825	**0.258**
PD3	-0.861	**-0.550**	**-0.067**	-0.895	**0.151**
PD4	-2.020	-2.155	**-0.045**	**-0.814**	-0.575
PD5	-2.198	-2.822	-0.245	-1.270	-0.472
AD2	-0.671	**0.368**	-0.001	-0.021	**0.512**
AD3	-0.897	-0.795	**0.084**	-0.178	-0.141
AD4	-2.125	-1.864	**0.191**	-0.374	-0.703
AD5	-2.304	-2.341	-0.075	**-0.088**	-1.182
CO2	-0.321	-0.083	-0.115	-0.126	-0.269
CO3	-0.640	**0.544**	-0.188	-0.637	**-0.077**
CO4	-1.379	-0.236	-0.300	**-0.607**	-1.876
CL2	-0.191	-0.488	-0.103	-0.110	-0.205
CL3	-0.416	**0.422**	**0.004**	-0.692	-0.581
CL4	-0.637	**0.027**	-0.168	**-0.439**	**-0.525**
FD2	**0.124**	-0.502	-0.138	-0.295	-0.112
FD3	-0.074	-0.587	-0.202	**-0.240**	-0.744
FD4	-0.429	-0.302	-0.330	-0.777	**-0.548**
SA2	-0.128	**0.624**	-0.020	**0.111**	-0.577
SA3	-0.404	**0.131**	-0.097	-0.093	-0.649
SA4	-0.605	-0.371	**-0.038**	-0.450	-1.093
SP2	**0.204**	-0.050	**0.110**	-0.027	-0.333
SP3	**0.006**	**0.323**	-0.048	**0.004**	**-0.314**
SP4	-0.269	-0.408	**-0.002**	**0.128**	-0.646
OC2	-0.093	-0.303	-0.011	-0.286	-0.238
OC3	-0.303	**-0.147**	**0.060**	-0.403	**-0.155**
OC4	**-0.163**	**-0.229**	-0.101	**-0.383**	-1.046
AC2	-0.140	**0.799**	-0.044	**0.271**	**0.039**
AC3	-0.494	**0.119**	-0.091	**0.325**	**0.095**
AC4	-0.582	**0.358**	-0.244	**0.119**	-0.472
DI2	**0.324**	-1.322	**0.067**	**0.105**	-0.240
DI3	-0.179	**-0.834**	**0.062**	**0.314**	-0.369
DI4	-0.201	**-0.964**	**0.023**	**0.161**	-0.585
Demographics					
Age Cat (18-60 and 60+)	-0.803	-3.353	-1.775	-0.631	0
Gender	-0.215	-0.858	-1.112	-0.894	0
Has Long-term Condition	-0.019	0.116	-0.180	-0.618	0
Class Share	0.241	0.110	0.286	0.160	0.203
N Obs[a]			14,625		
LL[b]			-9,006		

[a] Number of observations; [b] Log-Likelihood; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity;

Table 84: Six class latent class model (Model 70)

Parameter	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
MO2	-0.679	**0.485**	**0.466**	-0.045	**0.104**	-0.718
MO3	**-0.496**	-0.280	-0.484	**0.107**	**0.066**	-1.923
MO4	-1.569	**-0.161**	**-0.482**	-0.091	**0.376**	-4.050
MO5	-1.688	**0.535**	-1.576	-0.213	**0.267**	-4.485
SC2	-0.039	**0.597**	-0.220	-0.091	**0.925**	-0.907
SC3	-0.108	-0.076	-0.235	-0.087	-0.257	-1.604
SC4	-0.636	**0.435**	-1.341	-0.087	**-0.232**	-3.606
SC5	-0.721	**0.122**	-1.914	**0.033**	**-0.022**	**-2.445**
UA2	-0.196	-0.104	**0.133**	**0.001**	**0.073**	-0.222
UA3	-0.384	**0.424**	**0.218**	**0.138**	**0.229**	-1.143
UA4	-1.189	-0.272	-0.432	**0.047**	-0.143	**-0.390**
UA5	**-1.038**	**-0.254**	**-0.405**	**0.123**	**-0.038**	**-0.334**
PD2	-0.989	-0.304	-0.079	**-0.237**	-0.113	**0.060**
PD3	-1.006	**-0.173**	**0.006**	-0.145	-0.258	-0.362
PD4	-2.243	-1.540	-0.942	-0.193	**-0.154**	**1.179**
PD5	-2.394	-2.101	**-0.920**	-0.338	**0.067**	-0.710
AD2	-0.447	**0.342**	**0.208**	**0.124**	-0.552	**1.030**
AD3	-0.545	-1.224	-0.304	**0.155**	-0.934	**0.799**
AD4	-1.003	-4.264	-0.799	**0.266**	-1.004	-0.287
AD5	-1.134	-4.944	-1.182	**0.060**	-1.484	**0.646**
CO2	-0.317	-0.557	-0.086	-0.080	-0.486	**0.223**
CO3	**-0.283**	-0.746	-0.150	-0.157	-1.178	-0.405
CO4	-0.760	-1.486	-1.581	**-0.133**	-2.635	**-0.056**
CL2	-0.192	**0.274**	-0.342	**-0.205**	**0.190**	-0.495
CL3	-0.205	**0.206**	-1.060	-0.067	-0.229	**0.032**
CL4	-0.561	**0.404**	**-0.599**	-0.079	-0.872	**-0.081**
FD2	-0.026	-0.267	-0.423	-0.054	**0.298**	-0.698
FD3	-0.193	**-0.007**	-1.010	-0.107	-0.680	**-0.192**
FD4	-0.255	-0.129	**-0.939**	-0.208	**-0.571**	-1.811
SA2	**0.186**	-0.136	-0.423	**0.162**	-1.015	**-1.024**
SA3	-0.125	-0.599	-0.508	**0.206**	**-0.863**	-0.269
SA4	-0.434	**-0.543**	-0.808	**0.134**	-1.586	-0.760
SP2	-0.209	**0.385**	-0.244	**0.140**	**0.437**	-0.050
SP3	**0.079**	**0.056**	-0.368	**0.145**	-0.020	-0.910
SP4	-0.242	-0.269	-0.940	**0.154**	**0.171**	**-0.255**
OC2	**0.049**	-0.033	-0.289	**0.083**	-0.308	**-0.842**
OC3	-0.096	**0.160**	**-0.261**	**0.016**	**-0.303**	-0.600
OC4	**0.167**	-0.465	-1.279	-0.090	-0.496	-0.746
AC2	**0.205**	**0.696**	**0.069**	-0.018	**0.165**	-0.499
AC3	-0.121	**0.291**	**0.089**	**0.035**	-0.714	**-0.322**
AC4	**0.088**	-0.168	-0.647	-0.072	**-0.576**	-0.998
DI2	**0.029**	**-0.858**	-0.126	**0.042**	0.214	**0.257**
DI3	-0.156	**-0.322**	-0.416	**0.071**	-0.321	**0.464**
DI4	-0.203	**-0.256**	-0.418	**0.132**	**-0.140**	-0.533
Demographic						
Age Cat (18-60 and 60+)	-0.498	-0.572	0.765	-1.621	0.122	0
Gender	0.681	0.819	1.427	0.049	0.736	0
Has Long-term Condition	0.318	0.309	0.830	0.212	0.142	0
Class Share	0.281	0.100	0.189	0.246	0.098	0.086
N Obs[a]			14,625			
LL[b]			-8,963			

[a] Number of observations; [b] Log-Likelihood; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity;

8.17. Appendix 17: Further exploratory mixed logit models

Table 85: Mixed Logit models – Combining EQ-5D-5L and ASCOT as random parameters (overall level)

	Model 71: Overall dimension (1)		Model 72: Overall dimension (2)		Model 73: Overall dimension (log-normal)		Model 74: Overall dimension (correlated)	
Parameter	**Coef.[a]**	**SD[b]**	**Coef.**	**SD**	**Coef.**	**SD**	**Coef.**	**SD**
MO	-0.271***	0.296***	-0.276***	0.316***	-1.815***	1.036***	-0.221***	0.282***
SC	-0.193***	0.254***	-0.204***	0.278***	-2.179***	1.036***	-0.166***	0.247***
UA	-0.132***	0.115***	-0.140***	0.161***	-2.255***	0.589***	-0.107***	0.149***
PD	-0.273***	0.272***	-0.294***	0.296***	-1.728***	1.080***	-0.232***	0.276***
AD	-0.275***	0.326***	-0.281***	0.349***	-2.072***	1.392***	-0.216***	0.319***
CO	-0.251***	0.316***	-0.264***	0.330***	-1.806***	1.057***	-0.234***	0.302***
CL	-0.124***	0.025	-0.130***	0.077	-2.181***	0.249	-0.110***	0.105***
FD	-0.151***	0.125***	-0.160**	0.140*	-2.284***	0.844***	-0.137***	0.173***
SA	-0.114***	0.161***	-0.121***	0.181***	-2.841***	1.127***	-0.104***	0.198***
SP	-0.062***	0.098**	-0.064***	0.147***	-3.507***	1.172***	-0.051***	0.130***
OC	-0.086***	0.075***	-0.088***	0.187***	-3.299***	1.331***	-0.076***	0.174***
AC	-0.092***	0.043	-0.095***	0.103	-2.392***	0.298	-0.086***	0.159***
DI	-0.086***	0.081*	-0.090***	0.126**	-2.888***	0.846***	-0.074***	0.182**
N obs[c]	14,625		14,625		14,625		14,625	
LL[d]	-8,818		-8,760		-8,799		-8,959	
AIC[e]	17,647		17,572		17,651		18,016	
BIC[f]	17,862		17,788		18,866		18,421	

[a] Coefficient estimate; [b] standard deviation; [c] number of observations; [d] Log-Likelihood; [e] Akaike Information Criterion; [f] Bayesian Information Criterion; p-values indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at 0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

Table 86: Further exploratory models with all parameters, EQ-5D-5L and ASCOT as random

	Model 75: All dimension levels		Model 76: EQ-5D-5L dimensions vary		Model 77: ASCOT dimensions vary	
Parameter	**Coef.[a]**	**SD[b]**	**Coef.**	**SD**	**Coef.**	**SD**
MO2	-0.176**	0.234	-0.123*	0.041	-0.140*	N/A
MO3	-0.344***	0.011	-0.267***	0.119	-0.282***	N/A
MO4	-0.834***	0.443***	-0.663***	0.386***	-0.683***	N/A
MO5	-1.099***	0.805***	-0.915***	0.686***	-0.890***	N/A
SC2	-0.013	0.053	-0.011	0.091	-0.010	N/A
SC3	-0.228***	0.499***	-0.218***	0.337***	-0.226***	N/A
SC4	-0.658***	0.469***	-0.537***	0.360***	-0.531***	N/A
SC5	-0.718***	0.816***	-0.590***	0.558***	-0.564***	N/A
UA2	-0.065	0.208	-0.038	0.203	-0.042	N/A
UA3	**-0.036**	0.290*	**-0.021**	0.057	**-0.037**	N/A
UA4	-0.398***	0.204	-0.315***	0.108	-0.318***	N/A
UA5	**-0.389*** **	0.295**	**-0.310*** **	0.055	**-0.309*** **	N/A
PD2	-0.367***	0.298	-0.303***	0.012	-0.304***	N/A
PD3	**-0.349*** **	0.040	**-0.285*** **	0.058	**-0.280*** **	N/A
PD4	-0.939***	0.650***	-0.778***	0.438***	-0.758***	N/A
PD5	-1.158***	0.766***	-0.958***	0.589***	-0.926***	N/A
AD2	-0.052	0.398***	-0.031	0.342***	-0.056	N/A
AD3	-0.288***	0.377**	-0.222***	0.136	-0.232***	N/A
AD4	-0.799***	0.782***	-0.629***	0.454***	-0.657***	N/A
AD5	-1.002***	0.821***	-0.773***	0.346**	-0.792***	N/A
CO2	-0.170**	0.106	-0.172***	N/A	-0.143**	0.028
CO3	-0.321***	0.347**	-0.274***	N/A	-0.255***	0.237*
CO4	-0.891***	0.791***	-0.728***	N/A	-0.748***	0.709***
CL2	-0.148**	0.096	-0.126*	N/A	-0.117*	0.024
CL3	-0.253***	0.324**	-0.216***	N/A	-0.205***	0.262*
CL4	-0.388***	0.295*	-0.320***	N/A	-0.314***	0.081
FD2	-0.153**	0.319**	-0.109*	N/A	-0.118*	0.072
FD3	-0.337***	0.235	-0.279***	N/A	-0.287***	0.173
FD4	-0.486***	0.311	-0.390***	N/A	-0.403***	0.336**
SA2	-0.077	0.222	-0.060	N/A	-0.063	0.276**
SA3	-0.164**	0.087	-0.134**	N/A	-0.140**	0.039
SA4	-0.451***	0.364***	-0.357***	N/A	-0.361***	0.187
SP2	-0.013	0.174	-0.011	N/A	-0.012	0.244*
SP3	-0.065	0.192	-0.051	N/A	-0.058	0.026
SP4	-0.216***	0.355***	-0.175***	N/A	-0.159***	0.245*
OC2	-0.113	0.262	-0.091	N/A	-0.091	0.133
OC3	-0.126*	0.292	**-0.089**	N/A	**-0.090**	0.321***
OC4	-0.317***	0.482***	-0.253***	N/A	-0.271***	0.463***
AC2	**0.033**	0.108	**0.036**	N/A	**0.028**	0.025
AC3	-0.078	0.368***	-0.056	N/A	-0.074	0.265**
AC4	-0.274***	0.285*	-0.217***	N/A	-0.232***	0.210
DI2	-0.015	0.295**	-0.008	N/A	-0.011	0.112
DI3	-0.080	0.069	-0.061	N/A	-0.077	0.043
DI4	-0.208***	0.437***	-0.170***	N/A	-0.175***	0.315***
N obs[c]	14,625		14,625		14,625	
LL[d]	-8,809		-8,944		-8,904	
AIC[e]	17,852		17,902		17,998	
BIC[f]	18,626		18,432		18,601	

[a] Coefficient estimate; [b] standard deviation; c number of observations; d Log-Likelihood; e Akaike

Information Criterion; f Bayesian Information Criterion; p-values indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at 0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

Table 87: Mixed logit models – Further exploratory combinations of parameters in random

	Model 78: Most severe levels random		**Model 79: Top 20 from M 23**		**Model 80: Top 10 from M 23**		**Model 81: Top 5 from M 23**	
Parameter	**Coef.[a]**	**SD[b]**	**Coef.**	**SD**	**Coef.**	**SD**	**Coef.**	**SD**
MO2	-0.155**	N/A	-0.159**	N/A	-0.154**	N/A	-0.149**	N/A
MO3	-0.293***	N/A	-0.316***	N/A	-0.297***	N/A	-0.282***	N/A
MO4	-0.705***	N/A	-0.765***	0.349**	-0.727***	N/A	-0.678***	N/A
MO5	-0.968***	0.754***	-1.032***	0.804***	-0.989***	0.750***	-0.926***	0.704***
SC2	-0.019	N/A	-0.020	N/A	-0.019	N/A	-0.010	N/A
SC3	-0.236***	N/A	-0.257**	0.392**	-0.249***	N/A	-0.224***	N/A
SC4	-0.570***	N/A	-0.614**	0.501***	-0.588***	0.464***	-0.543***	N/A
SC5	-0.633***	0.710***	-0.665***	0.747***	-0.641***	0.718***	-0.598***	0.674***
UA2	-0.052	N/A	-0.049	N/A	-0.054	N/A	-0.047	N/A
UA3	**-0.038**	N/A	**-0.034**	N/A	**-0.037**	N/A	**-0.036**	N/A
UA4	-0.342***	N/A	-0.362	N/A	-0.357***	N/A	-0.325***	N/A
UA5	-0.345***	0.242**	**-0.354**	N/A	**-0.343*** **	N/A	**-0.316*** **	N/A
PD2	-0.320***	N/A	-0.340***	0.241	-0.321***	N/A	-0.312***	N/A
PD3	**-0.305*** **	N/A	**-0.327*** **	N/A	**-0.307*** **	N/A	**-0.294*** **	N/A
PD4	-0.801***	N/A	-0.863***	0.537***	-0.826***	0.558***	-0.769***	N/A
PD5	-1.022***	0.694***	-1.084***	0.732***	-1.041***	0.706***	-0.962***	N/A
AD2	-0.063	N/A	-0.053	0.262	-0.057	N/A	-0.058	N/A
AD3	-0.254***	N/A	-0.267***	0.299	-0.256***	N/A	-0.239***	N/A
AD4	-0.685***	N/A	-0.745***	0.696***	-0.720***	0.706***	-0.676***	0.680***
AD5	-0.872***	0.691***	-0.933***	0.770***	-0.892***	0.772***	-0.841***	0.704***
CO2	-0.149**	N/A	-0.169***	N/A	-0.164***	N/A	-0.158***	N/A
CO3	-0.268***	N/A	-0.303***	N/A	-0.282***	N/A	-0.264***	N/A
CO4	-0.778***	0.745***	-0.830***	0.740***	-0.798***	0.730***	-0.743***	0.708***
CL2	-0.139**	N/A	-0.139**	N/A	-0.144**	N/A	-0.123**	N/A
CL3	-0.232***	N/A	-0.239***	0.274*	-0.241***	N/A	-0.229***	N/A
CL4	-0.344***	0.119	-0.364***	N/A	-0.358***	N/A	-0.338***	N/A
FD2	-0.121*	N/A	-0.128*	N/A	-0.122*	N/A	-0.120***	N/A
FD3	-0.296***	N/A	-0.311***	N/A	-0.293***	N/A	-0.276***	N/A
FD4	-0.427***	0.318*	-0.452***	0.299	-0.435***	0.338**	-0.402***	N/A
SA2	-0.060	N/A	-0.064	N/A	-0.060	N/A	-0.060	N/A
SA3	-0.137**	N/A	-0.146**	N/A	-0.139**	N/A	-0.133**	N/A
SA4	-0.380***	0.194	-0.398***	N/A	-0.387***	N/A	-0.372***	N/A
SP2	-0.024	N/A	-0.019	0.271*	-0.020	N/A	-0.029	N/A
SP3	-0.062	N/A	-0.069	N/A	-0.064	N/A	-0.070	N/A
SP4	-0.180***	0.291**	-0.196***	0.230	-0.189***	N/A	-0.174***	N/A
OC2	-0.097	N/A	-0.094	N/A	-0.091	N/A	-0.086	N/A
OC3	-0.110*	N/A	-0.096	0.331**	-0.101*	N/A	-0.091	N/A
OC4	-0.281***	0.493***	-0.291***	0.521***	-0.276***	0.510***	-0.263***	N/A
AC2	**0.025**	N/A	**0.032**	N/A	**0.031**	N/A	**0.033**	N/A
AC3	-0.074	N/A	-0.067	N/A	-0.071	N/A	-0.069	N/A
AC4	-0.242***	0.175	-0.250***	N/A	-0.241***	N/A	-0.218***	N/A
DI2	-0.012	N/A	-0.018	N/A	-0.015	N/A	-0.006	N/A

DI3	-0.075	N/A	-0.077	N/A	-0.073	N/A	-0.064	N/A
DI4	-0.185***	0.326***	-0.205***	0.330**	-0.188***	N/A	-0.176***	N/A
N obs[c]	14,625		14,625		14,625		14,625	
LL[d]	-8,853		-8,949		-8,332		-8,861	
AIC[e]	17,821		17,776		17,773		17,820	
BIC[f]	18,293		18,306		18,220		18,226	

a Coefficient estimate; b standard deviation; c number of observations; d Log-Likelihood; e Akaike Information Criterion; f Bayesian Information Criterion; p-values indicated by stars: ***significant at 0.001, ** significant at 0.01; *significant at 0.05;; MO: mobility; SC: self-care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression; CO: control; CL: cleanliness; FD: food and drink; SA: safety; SP: social participation; OC: occupation; AC: accommodation; DI: dignity

8.18. Appendix 18: Latent class demographic parameters (design comparison study)

Table 88: Latent class demographic parameters (overlap designs A to D)

	Design A		Design B		Design C		Design D	
Parameter	C1[a]	C2	C1	C2	C1	C2	C1	C2
Class proportion (%)	68.3	31.7	65.7	34.3	21.2	78.8	69.0	31.0
Gender (1=M; 2=F)	0.364	0	0.494	0	-0.314	0	1.041	0
Age cat (1=<65; 2=>=65)	0.934	0	1.100	0	-0.912	0	0.587	0
Condition	-0.140	0	-0.028	0	-0.279	0	-0.599	0
Constant	-0.870	0	-1.422	0	0.398	0	-1.199	0

Table 89: Latent class demographic parameters (overlap designs E to G)

	Design E		Design F		Design G	
Parameter	C1	C2	C1	C2	C1	C2
Class proportion (%)	80.5	19.5	60.0	40.0	53.2	46.8
Gender (1=M; 2=F)	0.573	0	-0.318	0	-0.471	0
Age cat (1=<65; 2=>=65)	1.534	0	0.386	0	-0.487	0
Condition	-0.104	0	-1.025	0	-0.061	0
Constant	-1.195	0	0.886	0	1.490	0

Table 90: Latent class demographic parameters (non-overlap designs H to I)

	Design H				Design I			
Parameter	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
Class proportion (%)	22.5	30.3	15.1	32.1	44.4	24.1	13.3	18.2
Gender (1=M; 2=F)	-2.012	-1.930	-0.066	0	-0.211	-1.110	-0.012	0
Age cat (1=<65; 2=>=65)	0.695	-3.468	0.582	0	1.529	0.173	1.833	0
Condition	-1.147	0.296	-0.491	0	0.225	0.473	0.168	0
Constant	2.226	6.685	-1.188	0	-0.725	1.525	-2.639	0

Table 91: Latent class demographic parameters (non-overlap designs J to K)

	Design J			Design K		
Parameter	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**
Class proportion (%)	36.5	37.2	26.4	44.2	36.6	19.2
Gender (1=M; 2=F)	-0.454	0.530	0	-0.083	-0.085	0
Age cat (1=<65; 2=>=65)	-0.154	0.075	0	-1.087	-0.960	0
Condition	-0.497	-0.137	0	0.417	-0.155	0
Constant	1.405	-0.525	0	2.171	2.122	0

Table 92: Latent class demographic parameters (non-overlap designs L to M)

	Design L			Design M				
Parameter	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C4**	**C5**
Class proportion (%)	45.9	33.7	20.4	29.9	9.1	24.3	29.7	7.0
Gender (1=M; 2=F)	0.606	-0.725	0	1.335	0.567	1.101	0.330	0
Age cat (1=<65; 2=>=65)	-1.179	-2.224	0	1.423	1.536	2.459	1.917	0
Condition	1.315	1.780	0	-1.193	-1.697	-0.635	-1.133	0
Constant	0.933	3.653	0	-1.486	-1.432	-2.973	-0.605	0

Table 93: Latent class demographic parameters (non-overlap designs N to O)

	Design N				Design O			
Parameter	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**	**C3**	**C4**
Class proportion (%)	16.0	17.0	46.0	21.0	20.5	18.3	35.2	26.0
Gender (1=M; 2=F)	0.113	0.123	0.612	0	1.032	0.286	-0.154	0
Age cat (1=<65; 2=>=65)	-0.448	-0.187	0.179	0	-1.592	-0.614	0.365	0
Condition	0.812	0.741	0.062	0	0.541	0.564	0.285	0
Constant	-0.297	-0.528	-0.402	0	-0.253	-0.302	-0.115	0

Table 94: Latent class demographic parameters (non-overlap designs P to S)

	Design P		Design Q			Design R				Design S	
Parameter	**C1**	**C2**	**C1**	**C2**	**C3**	**C1**	**C2**	**C3**	**C4**	**C1**	**C2**
Class proportion (%)	31.4	68.6	19.9	29.7	50.4	23.6	15.1	56.4	14.9	35.0	65.0
Gender (1=M; 2=F)	0.391	0	-0.335	-1.290	0	-0.010	-0.588	0.923	0	-1.137	0
Age cat (1=<65; 2=>=65)	-0.734	0	-0.009	-2.230	0	-1.662	-0.402	-0.862	0	-1.110	0
Condition	-0.053	0	0.024	0.204	0	-0.114	-0.719	-0.816	0	0.352	0
Constant	-0.438	0	-0.397	3.934	0	2.707	1.791	1.314	0	2.295	0

References

1. OECD. Health Expenditure. 2019; Available from: https://www.oecd.org/health/health-expenditure.htm.
3. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. Canberra: Australian Department of Health, 2015.
3. National Institute for Health and Care Excellence. Methods for health technology assessment. London: National Institute for Health and Care Excellence, 2013.
4. Canadian Agency for Drugs and Technologies in Health. Guidelines for the Economic Evaluation of Health Technologies: Canada. 3rd edition. Ottawa: Canadian Agency for Drugs and Technologies in Health, 2017.
5. College voor zorgverzekeringen. Guidance for outcomes research 'for the assessment of the costeffectiveness of in-patient medicines'. Netherlands College voor zorgverzekeringen, 2008.
6. Von Neumann J, Morgenstern O. Theory of games and economic behavior. Theory of games and economic behavior. Princeton: Princeton University Press. 1944.
7. Seixas BV. Welfarism and extra-welfarism: a critical overview. Cadernos de Saúde Pública. 2017; 33(8).
8. Coast J. Maximisation in extra-welfarism: A critique of the current position in health economics. Social Science and Medicine. 2009; 69(5): 786-92.
9. Brouwer WB, Culyer AJ, van Exel NJA, Rutten FFH. Welfarism vs. extra-welfarism. Journal of Health Economics. 2008; 27(2): 325-38.
10. Boadway R, Bruce J. Welfare Economics. Oxford: Blackwell, 1984.
11. Gyrd-Hansen D. Willingness to pay for a QALY: theoretical and methodological issues. Pharmacoeconomics. 2005; 23(5): 423-32.
12. Round J. Is a QALY still a QALY at the end of life? Journal of Health Economics. 2012; 31: 521–27.
13. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care interventions. 4th Edition. Oxford: Oxford University Press, 2015
14. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health services research. 1972; 7(2): 118-33.
15. Weinstein MC, Torrance G, McGuire A. QALYs: The Basics. Value in Health. 2009; 12: S5-S9.
16. WHO, Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19–22 June 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948. In Grad, Frank P. (2002). "The Preamble of the Constitution of the World Health Organization". Bulletin of the World Health Organization. 80 (12): 982, 1948.
17. Mayo N. ISOQOL Dictionary of Quality of Life and Health Outcomes Measurement. ISOQOL, 2015.
18. Leidy NK, Revicki DA, Geneste B. Recommendations for evaluating the validity of quality of life claims for labeling and promotion. Value in Health. 1999; 2(2): 113-27.
19. Osoba D. Lessons learned from measuring health-related quality of life in oncology. Journal of Clinical Oncology. 1994; 12(3): 608-16.
20. Devlin N, Williams A. Valuing quality of life: results for New Zealand health professionals. New Zealand Medical Journal. 1999; 112(1083): 68-71.

21. Sun S, Chen J, Kind P, Xu L, Zhang Y, Burstrom K. Experience-based VAS values for EQ-5D-3L health states in a national general population health survey in China. Quality of Life Research. 2015; 24(3): 693-703.
22. Burström K, Sun S, Gerdtham UG, Hanriksson M, Johannesson M, Levin LA, Zethraeus N. Swedish experience-based value sets for EQ-5D health states. Quality of Life Research. 2014; 23(2): 431-42.
23. Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. Medical Decision Making. 1997; 17(2): 208-16.
24. Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. Medical Decision Making. 2001; 21(1): 17-27.
25. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? Health Economics. 2006; 15(7): 653-64.
26. Kahneman D, Tversky A. The psychology of preferences. Scientific American. 1982; 246(1): 160-73.
27. Tversky A, Kahneman D. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty. 1992; 5(4): 297-323.
28. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets in the UK. Journal of Health Economics. 2006; 25(2): 334-46.
29. Dolan P. Modeling valuations for EuroQol health states. Medical Care. 1997; 35(11): 1095-108.
30. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Economics. 2006; 15(4): 393–402.
31. Devlin NJ, Tsuchiya A, Buckingham K, Tilling C. A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. Health Economics. 2011; 20(3): 348-61.
32. Attema AE, Brouwer WB. On the (not so) constant proportional trade-off in TTO. Quality of Life Research. 2010; 19(4): 489-97.
33. Pinto-Prades JL, Attema A, Sánchez-Martínez FI. Measuring Health Utility in Economics. Oxford: Oxford University Press, 2019.
34. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Economics. 2002; 11(5): 447-56.
35. Augestad LA, Stavem K, Kristiansen IS, Samuelson CH, Rand-Hendriksen K. Influenced from the start: anchoring bias in time trade-off valuations. Quality of Life Research. 2016; 25(9): 2179 - 91.
36. Lancaster KJ. A New Approach to Consumer Theory. The Journal of Political Economy. 1966; 74(2): 132-57.
37. McFadden D. Conditional logit analysis of qualitative choice behaviour. In: Zarembka P, ed. Frontiers in Econometrics. New York: Academic Press, 1974: 105-142.
38. Johnson FR, Lancsar E, Marshall D, Kilambi V, Muhlbacher A, Regier DA, Bresnahan BW, Kanninen B, Bridges JFP. Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. Value in Health. 2013; 16(1): 3-13.
39. Bridges JF, Hauber AB, MArshall D, Lloyd A, Prosser LA, Regier DA, Johnson FR, Mauskopf J. Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in Health. 2011; 14(4): 403-13.
40. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Economics. 2012; 21(2): 145-72.

41. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob E. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics. 2014; 32(9): 883-902.
42. Coast J, Al-Janabi H, Sutton EJ, Horrocks SA, Vosper AJ, Swancutt DR, Flynn TN. Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. Health Economics. 2012; 21(6): 730-41.
43. Mulhern BJ, Norman R, Shah K, Bansback N, Longworth L, Viney R. How Should Discrete Choice Experiments with Duration Choice Sets Be Presented for the Valuation of Health States? Medical Decision Making. 2018; 38(3): 306-18.
44. Jonker MF, Donkers B, de Bekker-Grob E, Stolk EA. The Effect of Level Overlap and Color Coding on Attribute Non-attendance in Discrete Choice Experiments. Value in Health. 2018; 21(7): 767-71.
45. Mulhern B, Norman R, Lorgelly P, Lancsar E, Ratcliffe J, Brazier J, Viney R. Is Dimension Order Important when Valuing Health States Using Discrete Choice Experiments Including Duration? Pharmacoeconomics. 2017; 35(4): 439-51.
46. Norman R, Kemmler G, Viney R, Pickard AS, Gamper E, Holzner B, Nerich V, King M. Order of Presentation of Dimensions Does Not Systematically Bias Utility Weights from a Discrete Choice Experiment. Value in Health. 2016; 19(8): 1033-38.
47. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value in Health. 2014; 17(4): 445-53.
48. Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate health state utility values. Journal of Health Economics. 2012; 31(1): 306-18.
49. Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value in Health. 2010; 13(8): 1005-13.
50. Viney R, Norman R, Brazier J, Cronin P, King MT, Ratcliffe J, Street D. An Australian discrete choice experiment to value EQ-5D health states. Health Economics. 2014; 23(6): 729-42.
51. Jonker MF, Attema AE, Donkers B, Stolk EA, Versteegh MM. Are Health State Valuations from the General Public Biased? A Test of Health State Reference Dependency Using Self-assessed Health and an Efficient Discrete Choice Experiment. Health Economics. 2017; 26(12): 1534-47.
52. Hakim Z, Pathak DS. Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modelling. Health Economics. 1999; 8(2): 103-16.
53. Rowen D, Brazier J, van Hout B. A Comparison of Methods for Converting DCE Values onto the Full Health-Dead QALY Scale. Medical Decision Making. 2015; 35(3): 328-40.
54. Ramos-Goni JM, Pinto-Prades Jl, Oppe M, Cabases JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. Medical Care. 2017; 55(7): e51-e58.
55. Bansback N, Hole AR, Mulhern B, Tsuchiya A. Testing a discrete choice experiment including duration to value health states for large descriptive systems: addressing design and sampling issues. Social Science and Medicine. 2014; 114: 38-48.
56. Mulhern B, Bansback N, Brazier J, Buckingham K, Cairns J, Devlin N, Dolan P, Hole AR, Kavetsos G, Longworth L, Rowen D, Tsuchiya A. Preparatory study for the revaluation of the EQ-5D tariff: methodology report. Health Technology Assessment. 2014; 18(12): 1-191.
57. Norman R, Cronin P, Viney R. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. Applied Health Economics and Health Policy. 2013; 11(3): 287-98.

58. Norman R, Viney R, Burgess L, Cronin P, King M, Ratcliffe J, Street D. Valuing SF-6D Health States Using a Discrete Choice Experiment. Medical Decision Making. 2013; 34(6): 773-86.
59. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using Discrete Choice Experiments with Duration to Model EQ-5D-5L Health State Preferences. Medical Decision Making, 2017; 37(3): 285-97.
60. Jonker MF, Donkers B, de Bekker-Grob E, Stolk EA. Advocating a Paradigm Shift in Health-State Valuations: The Estimation of Time-Preference Corrected QALY Tariffs. Value in Health. 2018; 21(8): 993-1001.
61. Street DJ, Viney R. Design of Discrete Choice Experiments, in Oxford Encyclopedia of Health Economics. Oxford: Oxford University Press, 2019.
62. Street DJ, Burgess L. The Construction of Optimal Stated Choice Experiments. Hoboken: John Wiley & Sons, 2007.
63. Hedayat AS, Sloane NJA, Stufken J. Orthogonal Arrays: Theory and Applications. Rotterdam: Springer, 1999.
64. Huber J, Zwerina K. The Importance of Utility Balance in Efficient Choice Designs. Journal of Marketing Research. 1996; 33(3): 307-317.
65. Kuhfeld WF. Marketing Research Methods in SAS: Technical report. 2010; Available from: http://support.sas.com/resources/papers/tnote/tnotemarketresearch.html.
66. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LP, 2017.
67. ChoiceMetrics. Ngene [software for experimental design]. Sydney: Choice Metrics, 2014.
68. Johnson FR, Yang JC, Reed SC. The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments. Value in Health. 2019; 22(2): 157-60.
69. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Economics, Policy and Law. 2009; 4(4): 527-46.
70. Cheng S, Long JS. Testing for IIA in the Multinomial Logit Model. Sociological Methods and Research. 2007; 35(4): 583-600.
71. Swait J, Louviere J. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. Journal of Marketing Research. 1993; 30(3): 305-14.
72. Train KE. EM Algorithms for nonparametric estimation of mixing distributions. Journal of Choice Modelling. 2008; 1(1): p. 40-69.
73. Hensher DA, Greene WA. The Mixed Logit Model: The State of Practice. University of Sydney: Working Paper, 2002
74. Hess S, Train K. Correlation and scale in mixed logit models. Journal of Choice Modelling. 2017; 23: 1-8.
75. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Medical Care. 1992; 30(6): 473-83.
76. Fayers PMD. Quality of Life: The Assessment, Analysis and Reporting of Patient-reported Outcomes. 3rd Edition. London: Wiley Blackwell, 2016.
77. Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37(1): 53-72.
78. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of Life Research. 2011; 20(10): 1727-36.
79. Brazier JE, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics. 2002; 21(2): 271-92.
80. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Medical Care. 2004; 42(9): 851-9.

81. Richardson J, Iezzi A, Khan M, Maxwell A. Validity and Reliability of the Assessment of Quality of Life (AQoL)-8D Multi-Attribute Utility Instrument. Patient. 2014; 7(1): 85-96.
82. Richardson J, Sinha K, Iezzi A, Khan MA. Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. Quality of Life Research. 2014; 23(8): 2395-404.
83. Scuffham PA, Whitty JA, Mitchell A, Viney R. The use of QALY weights for QALY calculations: a review of industry submissions requesting listing on the Australian Pharmaceutical Benefits Scheme 2002-4. Pharmacoeconomics, 2008. 26(4): 297-310.
84. ISPOR, Pharmacoeconomic Guidelines Around the World. ISPOR Connections, 2016.
85. Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, Barkham M, Ibbotson R. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technology Assessment. 2012; 16(32): 1-114.
86. Stamuli E. Health outcomes in economic evaluation: who should value health? British Medical Bulletin. 2011; 97: 197-210.
87. Rowen D, Mulhern B, Banerjee S, Tait R, Watchurst C, Smith SC, Young TA, Knapp M, Brazier JE. Comparison of general population, patient, and carer utility values for dementia health states. Medical Decision Making. 2015; 35(1): 68-80.
88. Gandhi M, Tan RS, Ng R, Choo SP, Chia WK, Toh CK, Lam C, Lee PT, Latt NKZ, Rand-Hendriksen K, Cheung YB, Luo N. Comparison of health state values derived from patients and individuals from the general population. Quality of Life Research, 2017. 26(12): 3353-63.
89. Mulhern B, Rowen D, Snape D, Jacoby A, Marson T, Hughes D, Baker G, Brazier J. Valuations of epilepsy-specific health states: a comparison of patients with epilepsy and the general population. Epilepsy and Behaviour. 2014; 36: 12-7.
90. Szende A, Oppe M, Devlin N. EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Rotterdam: EuroQol Group Monographs, 2007.
91. Viney R, Norman R, King MT, Cronin P, Street DJ, Knox S, Ratcliffe J. Time trade-off derived EQ-5D weights for Australia. Value in Health. 2011; 14(6): 928-36.
92. Feng Y, Devlin NJ, Shah KK, Mulhern BJ, van Hout B. New methods for modelling EQ-5D-5L value sets: An application to English data. Health Economics. 2018; 27(1): 23-38.
94. Devlin, NJ, Shah K, Feng Y, Mulhern BJ, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. 2018; 27(1): 7-22.
95. Flattery M, Mulhern B, Norman R, Viney R, Street D, Feng Y, Addo R, Manipis K, Meshcheriakova E, Saing S., Valuing EQ-5D-5L in Australia using an adapted EQ-VT: Informing the further development of a revised valuation protocol. EuroQol Plenary, Barcelona, 2017.
96. Abellan Perpinan JM, Sanchez Martinez FI, Martinez Perez, Mendez I. Lowering the 'floor' of the SF-6D scoring algorithm using a lottery equivalent method. Health Economics. 2012; 21(11): 1271-85.
97. Brazier JE, Fukuhara S, Roberts J, Kharroubi S, Yamamoto Y, Ikeda S, Doherty J, Kurokawa K. Estimating a preference-based index from the Japanese SF-36. Journal of Clinical Epidemiology. 2009; 62(12): 1323-31.
98. Cruz LN, Camey SA, Hoffmann JF, Rowen D, Brazier JE, Fleck MP, Polanczyk CA. Estimating the SF-6D value set for a population-based sample of Brazilians. Value in Health. 2011; 14(5): S108-14.
99. Ferreira LN, Ferreira PL, Pereira LN, Brazier JE, Rowen D. A Portuguese value set for the SF-6D. Value in Health. 2010; 13(5): 624-30.
100. Lam CL, Brazier JE, McGhee SM. Valuation of the SF-6D Health States Is Feasible, Acceptable, Reliable, and Valid in a Chinese Population. Value in Health. 2008; 11(2): 295-303.

101. McGhee SM, Brazier J, Lam CLK, Wong LC, Chau J, Cheung A, Ho A. Quality-adjusted life years: population-specific measurement of the quality component. Hong Kong Medical Journal. 2011; 17(6): 17-21.
102. McCabe C, Brazier J, Gilks P, Tsuchiya A, Roberts J, O'Hagan A, Stevens K. Using rank data to estimate health state utility models. Journal of Health Economics. 2006; 25(3): 418-31.
103. Kharroubi SA, Brazier JE, Roberts J, O'Hagan A. Modelling SF-6D health state preference data using a nonparametric Bayesian method. Journal of Health Economics. 2007; 26(3): 597-612.
104. Furlong WJ, Feeny D, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. Annals of Medicine. 2001; 33(5): 375-84.
105. Horsman J, Furlong W, Feeny D, Torrance GW. The Health Utilities Index (HUI®): concepts, measurement properties and applications. Health and Quality of Life Outcomes. 2003; 1(54).
106. Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzmann S. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. Journal of Clinical Oncology. 1992; 10(6): 923-8.
107. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M, Boyle M. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Medical Care. 2002; 40(2): 113-28.
108. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of Health-Related Quality of Life. Quality of Life Research. 1999; 8: 209-24.
109. Longworth L, Yang Y, Young T, Mulhern B, Hernandez Alava M, Mukuria C, Rowen D, Tosh J, Tsuchiya A, Evans P, Keetharuth AD, Brazier J. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. Health Technology Assessment. 2014; 18(9): 1-224.
110. Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in schizophrenia? A systematic review. Value in Health. 2011; 14(6): 907-20.
111. Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, Lloyd Jones M, Paisley S, O'Cathain A, Barkham M, Knapp M, Byford S, Gilbody S, Parry G. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technology Assessment. 2014; 18(34): 1-188.
112. Connell J, O'Cathain A, Brazier J. Measuring quality of life in mental health: Are we asking the right questions? Social Science and Medicine. 2014; 120: 12-20.
113. Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with Type 2 diabetes mellitus. Diabetic Medicine. 2011; 28(4): 395-413.
114. Adams R, Walsh C, Veale D, Bresnihan B, Fitzgerald O, Barry M. Understanding the relationship between the EQ-5D, SF-6D, HAQ and disease activity in inflammatory arthritis. Pharmacoeconomics. 2010. 28(6): 477-87.
115. Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S. Psychometric Comparison of the Standard EQ-5D to a 5 Level Version in Cancer Patients. Medical Care. 2007; 45(3): 259-263.

116. Brazier J, Tsuchiya A. Improving Cross-Sector Comparisons: Going Beyond the Health-Related QALY. Applied Health Economics and Health Policy. 2015; 13(6): 557-65.
117. Brazier JE, Rowen D, Lloyd A, Karimi M. Future Directions in Valuing Benefits for Estimating QALYs: Is Time Up for the EQ-5D? Value in Health. 2019; 22(1): 62-68.
118. Neumann PJ, Cohen JT. QALYs in 2018—Advantages and Concerns. Journal of the American Medical Association. 2018; 319(24): 2473-74.
119. Pettitt DA, Raza S, Naughton B, Roscoe A, Ramakrishnan A, All A, Davies B, Dopson S, Hollander G, Smith JA, Brindley DA. The limitations of QALY: a literature review. Journal of Stem Cell Research and Therapy. 2016; 6(4).
120. Soares MO. Is the QALY blind, deaf and dumb to equity? NICE's considerations over equity. British Medical Bulletin. 2012; 101(1): 17-31.
121. Goldstein DA. Using Quality-Adjusted Life-Years in Cost-Effectiveness Analyses: Do Not Throw Out the Baby or the Bathwater. Journal of Oncology Practice. 2016; 12(6): 500-502.
122. Kahneman D. A Different Approach to Health State Valuation. Value in Health. 2009; 12(S1): S16-S17.
123. Netten A, Burge P, Malley J, Potoglou D, Towers AM, Brazier J, Flynn T, Forder J, Wall B. Outcomes of social care for adults: developing a preference-weighted measure. Health Technology Assessment. 2012; 16(16): 1-166.
124. Al-Janabi H, Flynn T, Coast J. Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. Quality Life Research. 2012; 21(1): 167-76.
125. Mulhern B, Norman R, Street DJ, Viney R. One Method, Many Methodological Choices: A Structured Review of Discrete-Choice Experiments for Health State Valuation. Pharmacoeconomics. 2019; 37(1): 29-43.
126. Cheung KL, Wijnen BFM, Hollin IL, Janssen EM, Bridges JF, Evers SMAA, Hiligsmann M. Using Best-Worst Scaling to Investigate Preferences in Health Care. Pharmacoeconomics. 2016; 34(12): 1195-209.
127. Karimi M, Brazier J, Paisley S. How do individuals value health states? A qualitative investigation. Social Science and Medicine. 2017; 172: 80-88.
128. Kim SH, Ahn J, Ock M, Shin S, Park J, Luo N, Jo MW. The EQ-5D-5L valuation study in Korea. Quality of Life Research. 2016; 25(7): 1845-52.
129. Xie F, Pickard AS, Krabbe PFM, Revicki D, Viney R, Devlin N, Feeny D. A Checklist for Reporting Valuation Studies of Multi-Attribute Utility-Based Instruments (CREATE). Pharmacoeconomics. 2015; 33(8): 867-77.
130. Xie F, Pullenayegum E, Gaebel K, Oppe M, Krabbe PFM. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best-worst scaling? European Journal of Health Economics. 2014; 15(3): 281-8.
131. Scalone L, Stalmeier PFM, Milani S, Krabbe PFM. Values for health states with different life durations. European Journal of Health Economics. 2014; 16: 917-25.
132. Ryan M, Netten A, Skatun D, Smith P. Using discrete choice experiments to estimate a preference-based measure of outcome--an application to social care for older people. Journal of Health Economics. 2006; 25(5): 927-44.
133. Rowen D, Mulhern B, Stevens K, Vermaire JH. Estimating a Dutch Value Set for the Paediatric Preference-Based CHU9D Using a Discrete Choice Experiment with Duration. Value in Health. 2018; 21(10): 1234-42.
134. Robinson A, Spencer A, Moffatt P. A Framework for Estimating Health State Utility Values within a Discrete Choice Experiment: Modeling Risky Choices. Medical Decision Making. 2014; 35(3): 341-50.

135. Ratcliffe J, Brazier J, Tsuchiya A, Symonds T, Brown M. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Economics. 2009; 18(11): 1261-76.
136. Ramos-Goni JM, Rivero-Arias O, Errea M, Stolk EA, Herdman M, Cabases JM. Dealing with the health state 'dead' when using discrete choice experiments to obtain values for EQ-5D-5L heath states. European Journal of Health Economics. 2013; 14(S1): S33-42.
137. Ramos-Goñi JM, Craig BM, Oppe M, Ramallo-Farina Y, Pinto-Prades JL, Luo N, Rivero-Arias O. Handling Data Quality Issues to Estimate the Spanish EQ-5D-5L Value Set Using a Hybrid Interval Regression Approach. Value in Health. 2018; 21(5): 596-604.
138. Purba FD, Hunfeld JAM, Timman R, Sadarjoen SS, Passchier J, Busschbach JJV. Test-Retest Reliability of EQ-5D-5L Valuation Techniques: The Composite Time Trade-Off and Discrete Choice Experiments. Value in Health. 2018; 21(2): 1243-49.
139. Purba FD, Hunfeld JAM, Iskandersyah A, Fitriana TS, Sadarjoen SS, Ramos-Goni JM, Passchier J, Busschbach JJV. The Indonesian EQ-5D-5L Value Set. Pharmacoeconomics, 2017; 35(11): 1153-65.
140. Pullenayegum, E. and F. Xie, Scoring the 5-level EQ-5D: can latent utilities derived from a discrete choice model be transformed to health utilities derived from time tradeoff tasks? Med Decis Making, 2013. 33(4): p. 567-78.
141. Potoglou D, Burge P, Flynn T, Netten A, Malley J, Forder J, Brazier JE. Best-worst scaling vs. discrete choice experiments: an empirical comparison using social care data. Social Science and Medicine. 2011; 72(10): 1717-27.
142. Krucien N, Watson V, Ryan M. Is Best-Worst Scaling Suitable for Health State Valuation? A Comparison with Discrete Choice Experiments. Health Economics, 2017; 26(12): e1-e16.
143. Krabbe PFM, Devlin NJ, Stolk EA, Shah K, Oppe M, van Hout B, Quik EH, Pickard AS, Xie F. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. Medical Care. 2014; 52(11): 935-43.
144. King M, Viney R, Pickard AS, Rowen D, Aaronson NK, Brazier JE, Cella D, Costa DSJ, Fayers PM, Kemmler G, McTaggart-Cowen H, Mercieca-Bebber R, Peacock S, Street DJ, Young TA, Norman R. Australian Utility Weights for the EORTC QLU-C10D, a Multi-Attribute Utility Instrument Derived from the Cancer-Specific Quality of Life Questionnaire, EORTC QLQ-C30. Pharmacoeconomics. 2018; 36(2): 225-38.
145. Jakubczyk M, Craig BM, Barra M, Groothuis-Oudshoorn C, Hartman J, Huynh E, Ramos-Goñi JM, Stolk EA, Rand K. Choice Defines Value: A Predictive Modeling Competition in Health Preference Research. Value in Health. 2018; 21(2): 229-38.
146. Huynh E, Coast J, Rose J, Kinghorn P, Flynn T. Values for the ICECAP-Supportive Care Measure (ICECAP-SCM) for use in economic evaluation at end of life. Social Science and Medicine. 2017; 189: 114-28.
147. Hole AR, Norman R, Viney R. Response patterns in health state valuation using endogenous attribute attendance and latent class analysis. Health Economics. 2016; 25(2): 212-24.
148. Hoefman RJ, van Exel J, Rose JM, van de Wetering EJ, Brouwer WBF. A discrete choice experiment to obtain a tariff for valuing informal care situations measured with the CarerQol instrument. Medical Decision Making. 2014; 34(1): 84-96.
149. Hoefman RJ, van Exel J, Brouwer WBF. Measuring Care-Related Quality of Life of Caregivers for Use in Economic Evaluations: CarerQol Tariffs for Australia, Germany, Sweden, UK, and US. Pharmacoeconomics. 2017; 35(4): 469-78.

150. Hauber AB, Mohamed AF, Johnson FR, Oyelowo O, Curtis BH, Coon C. Estimating importance weights for the IWQOL-Lite using conjoint analysis. Quality of Life Research. 2010; 19(5): 701-9.
151. Gu Y, Norman R, Viney R. Estimating health state utility values from discrete choice experiments--a QALY space model approach. Health Economics. 2014; 23(9): 1098-114.
152. Goossens LMA, Rutten-van Molken MPMH, Boland MRS, Donkers B, Jonker M, Slok AHM, Salome PL, van Schayck OCP, Veen JCCM, Stolk EA. ABC Index: quantifying experienced burden of COPD in a discrete choice experiment and predicting costs. BMJ Open. 2017; 7(12): e017831.
153. Gamper EM, Holzner B, King MT, Norman R, Viney R, Nerich V, Kemmler G. Test-Retest Reliability of Discrete Choice Experiment for Valuations of QLU-C10D Health States. Value in Health. 2018; 21(8): 958-66.
154. Feng Y, Hole AR, Karimi M, Tsuchiya A, van Hout B. An exploration of the non-iterative time trade-off method to value health states. Health Economics. 2018; 27(8): 1248-63.
155. Cole A, Shah K, Mulhern B, Feng Y, Devlin N. Valuing EQ-5D-5L health states 'in context' using a discrete choice experiment. Health Economics. 2018; 19(4): 595-605.
156. Burr JM, Kilonzo M, Vale L, Ryan M. Developing a preference-based Glaucoma Utility Index using a discrete choice experiment. Optometry and Vision Science. 2007; 84(8): 797-808.
157. Bailey H. Results from a preliminary study to develop the quality adjustments for quality adjusted life year values for Trinidad and Tobago. West Indian Medical Journal. 2013; 62(6): 543-7.
158. Craig BM, Pickard AS, Stolk E, Brazier JE. US valuation of the SF-6D. Medical Decision Making. 2013; 33(6): 793-803.
159. van Hoorn RA, Donders ART, Oppe M, Stalmeier PFM. The Better than Dead Method: Feasibility and Interpretation of a Valuation Study. Pharmacoeconomics. 2014; 32(8): 789-99.
160. Craig BM, Reeve BB, Brown PM, Cella D, Hays RD, Lipscomb J, Pickard AS, Revicki DA. US valuation of health outcomes measured using the PROMIS-29. Value in Health. 2014; 17(8): 846-53.
161. Gärtner FR, de Bekker-Grob E, Stiggelbout A, Rijnders ME, Freeman LM, Middeldorp J, Bloemenkamp K, de Miranda E, van den Akker-van Marle ME. Calculating Preference Weights for the Labor and Delivery Index: A Discrete Choice Experiment on Women's Birth Experiences. Value in Health. 2015; 18(6): 856-64.
162. Mulhern B, Shah K, Janssen MF, Longworth L, Ibbotson R. Valuing Health Using Time Trade-Off and Discrete Choice Experiment Methods: Does Dimension Order Impact on Health State Values? Value in Health. 2016; 19(2): 210-7.
163. Norman R, Mulhern B, Viney, R. The Impact of Different DCE-Based Approaches When Anchoring Utility Scores. Pharmacoeconomics. 2016; 34(8): 805-14.
164. Shiroiwa T, Ikeda S, Noto S, Igarashu A, Fukuda T, Saito S, Shimozuma K. Comparison of Value Set Based on DCE and/or TTO Data: Scoring for EQ-5D-5L Health States in Japan. Value in Health. 2016; 19(5): 648-54.
165. Norman R, Viney R, Aaronson NK, Brazier JE, Costa DSJ, Fayers PM, Kemmler G et al. Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format. Quality of Life Research. 2016; 25(3): 637-49.
166. Craig B, Greiner W, Brown D, Reeve B. Valuation of Child Health-Related Quality of Life in the United States. Health Economics. 2016; 25(6): 768-77.

167. Craig BM, Brown DS, Reeve B. Valuation of Child Behavioral Problems from the Perspective of US Adults. Medical Decision Making. 2016; 36(2): 199-209.
168. Versteegh M, Vermeulen K, Evers MAA, de Wit A, Prenger R, Stolk E. Dutch Tariff for the Five-Level Version of EQ-5D. Value in Health. 2016; 19(4): 343-52.
169. Bailey H, Stolk E, Kind P. Toward Explicit Prioritization for the Caribbean: An EQ-5D Value Set for Trinidad and Tobago. Value in Health Regional Issues. 2016; 11: 60-7.
170. Robinson A, Spencer AE, Pinto-Prades JL, Covey JA. Exploring Differences between TTO and DCE in the Valuation of Health States. Medical Decision Making, 2017; 37(3): 273-84.
171. Xie F, Pullenayegum E, Pickard AS, Ramos-Goni JM, Jo MW, Igarashi A. Transforming Latent Utilities to Health Utilities: East Does Not Meet West. Health Economics. 2017; 26(12): 1524-33.
172. Craig BM, Rand K. Choice Defines QALYs: A US Valuation of the EQ-5D-5L. Medical Care. 2018; 56(6): 529-36.
173. Craig BM, Rand K, Stalmeier PFM. Quality-Adjusted Life-Years without Constant Proportionality. Value in Health. 2018; 21(9): 1124-31.
174. Hole AR. CLOGITHET: Stata Module to Estimate Heteroscedastic Conditional Logit Models. Statistical Software Components, 2006.
175. Hole AR. Small-sample properties of tests for heteroscedasticity in the conditional logit model. Economics Bulletin. 2006; 3(18): 1 - 14.
176. Edelen, M.O. and B.B. Reeve, Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of LIfe Research. 2007; 16(S1): 5-18.
177. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research , 1960/1980.
178. Samejima F. Graded Response Model, in Handbook of Modern Item Response Theory, W.J. cvan der Linden and R.K. Hambleton, Editors. Springer: New York, 1997.
179. Embretson S, Reise SP. The new rules of measurement., in Item response theory for psychologists. Mahwah: Lawrence Erlbaum, 2000.
180. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Medical Care. 2000; 38(S9): II28-42.
181. Smith W, Patel A, McCrone P, Jin H, Osumili B, Barrett B. Reducing outcome measures in mental health: a systematic review of the methods. Journal of Mental Health. 2016; 25(5): 461-72.
182. Watt T, Bjorner JB, Groenvold M, Cramon P, Hillert K, Hegedus L et al. Development of a Short Version of the Thyroid-Related Patient-Reported Outcome ThyPRO. Thyroid. 2015; 25(10): 1069-79.
183. Xia J, Tang Z, Wu P, Wang J, Yu J. Use of item response theory to develop a shortened version of the EORTC QLQ-BR23 scales. Scientific Reports. 2019; 9(1): 1764.
184. van Nispen RM, Knol DL, Langelaan M, de Boer MR, Terwee CB, van Rens GHMB. Applying multilevel item response theory to vision-related quality of life in Dutch visually impaired elderly. Optometry and Visoon Science. 2007; 84(8): 710-20.
185. van Nispen RM, Knol DL, Langelaan M, van Rens GHMB. Re-evaluating a vision-related quality of life questionnaire with item response theory (IRT) and differential item functioning (DIF) analyses. BMC Medical Research Methodology. 2011; 11: 125.
186. Chamot E, Kister I, Cutter G. Item response theory-based measure of global disability in multiple sclerosis derived from the Performance Scales and related items. BMC Neurology. 2014; 14: 192.
187. Luo Y, Yang J, Zhang Y. Development and validation of a patient-reported outcome measure for stroke patients. Health and Quality of Life Outcomes. 2015; 13: 53.

188. Bjorner JB, Petersen MA, Groenvold M, Aaronson N, Ahlner-Elmqvist M, Arraras JI, Bredart A, Fayers P, Jordhoy M, Sprangers M, Watson M, Young T. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. Quality of Life Research 2004; 13(10): 1683-97.
189. Petersen MA, Groenvold M, Aaronson N, Sprangers M, Bjorner JB. Multidimensional Computerized Adaptive Testing of the EORTC QLQ-C30: Basic Developments and Evaluations. Quality of Life Research.2006; 15(3): 315-29.
190. Reeve BB, Hays R, Bjorner JB, Cook K, Crane P, Teresi J, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care. 2007; 45(5 Suppl1): S22-31.
191. Goodwin E, Green C. A Systematic Review of the Literature on the Development of Condition-Specific Preference-Based Measures of Health. Applied Health Economics and Health Policy. 2016;14(2): 161-83.
192. Rowen D, Brazier J, Young T, Gaugris S, King MT, Craig B, Velikova G. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. Value in Health. 2011; 14(5): 721-31.
193. Mulhern B, Rowen D, Brazier J, Smith S, Romeo R, Tait R, Watchurst C, Chua K-C, Loftus V, Young T, Lamping D, Knapp M, Howard R, Banerjee S. Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation. Health Technology Assessment. 2013; 17(5): 1-140.
194. Keetharuth AD, Brazier J, Connell J, Bjorner JB, Carlton J, Taylor Buck E et al. Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. British Journal of Psychiatry. 2018; 212(1): 42-49.
195. Izumi R, Noto S, Takamoto U, Ikeda S, Fukuda T. Comparison of three utility measures using item response theory in stroke patients. Quality of Life Research. 2013; 22.
196. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim JS. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. Medical decision making. 2010; 30(1): 5-15.
197. Petrillo J, Cana SJ, McLeod L, Coon C. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. Value in Health. 2015; 18(1): 25-34.
198. Stover AM, Mc Leod LD, Langer MM, Chen WH, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. Journal of Patient-Reported Outcomes. 2019; 3(1): 50.
199. Nolte S, Coon C, Hudgens S, Verdam MGE. Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods. Journal of Patient-Reported Outcomes. 2019; 3(1): 46.
200. Cleanthous S, Barbic SP, Smith S, Regnault A. Psychometric performance of the PROMIS® depression item bank: a comparison of the 28- and 51-item versions using Rasch measurement theory. Journal of Patient-Reported Outcomes. 2019; 3(1): 47.
201. Bjorner JB. State of the psychometric methods: comments on the ISOQOL SIG psychometric papers. Journal of Patient-Reported Outcomes. 2019; 3(1): 49.
202. Cappelleri JC, Lundy J, Hays R. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clinical Therapeutics. 2014; 36(5): 648-62.
203. Ware JE. SF-36 Health Survey Update. Spine. 2000; 25(24): 3130-39.

204. Cella D, Riley W, Stone A, Rothrock N, Reeve BB, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. Journal of Clinical Epidemiology. 2010; 63(11): 1179-94.
205. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health and Quality of Life Outcomes. 2007; 5(1): 63.
206. Office of National Statistics Personal Well-being in the UK. Newport: Office for National Statistics, 2016.
207. Richardson J, Khan M, Iezzi A, Maxwell A. Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D multi-attribute utility instruments. Medical Decision Making. 2015; 35(3): 276-91.
208. PROMIS, PROMIS ADULT PROFILE INSTRUMENTS: A brief guide to the PROMIS® Profile instruments for adult respondents. Patient Reported Outcomes Measurement Information System, 2019.
209. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics. 2004; 13(9): 873-84.
210. Yong AG, Pearce S. A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutorials in Quantitative Methods for Psychology. 2013; 9: 79-94.
211. Cai L. High-dimensional Exploratory Item Factor Analysis by A Metropolis–Hastings Robbins–Monro Algorithm. Psychometrika. 2010; 75(1): 33-57.
212. Cai L. Metropolis–Hastings Robbins–Monro Algorithm for Confirmatory Item Factor Analysis. Journal of Educational and Behavioral Statistics 2010; 35(3): 307-35.
213. Tabachnick BG, Fidell LS. Using Multivariate Statistics. New York: Allyn and Bacon, 2007.
214. Costello AB, Osborne JW. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research, & Evaluation. 2005; 10: 1-9.
215. Hilari K, Byng S, Lamping D, Smith SC. Stroke and Aphasia Quality of Life Scale-39 (SAQOL-39): evaluation of acceptability, reliability, and validity. Stroke. 2003; 34(8): 1944-50.
216. Browne M. An overview of analytic rotation in exploratory factor analysis. Multivariate Behavioural Research. 2001; 36: 111-50.
217. Paek I, Han KT. IRTPRO 2.1 for Windows (Item Response Theory for Patient-Reported Outcomes). Applied Psychological Measurement. 2012; 37(3): 242-52.
218. Martin M, Kosinski M, Bjorner JB, Ware JE, Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Quality of lIfe Research. 2007; 16(4): 647-60.
219. Toland MD. Practical Guide to Conducting an Item Response Theory Analysis. The Journal of Early Adolescence. 2013; 34(1): 120-51.
220. Orlando M, Thissen D. Further Investigation of the Performance of S-X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. Applied Psychological Measurement. 2003; 27(4): 289-98.
221. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19: 716-23.
222. Stone M. Comments on model selection criteria of Akaike and Schwarz. (Series B).Journal of the Royal Statistical Society. 1979; 41: 276-78.

223. Wildman J, McMeekin P, Grieve E, Briggs A. Economic evaluation of integrated new technologies for health and social care: Suggestions for policy makers, users and evaluators. Social Science and Medicine. 2016; 169: 41-48.
224. van Leeuwen KM, Bosmans JE, Jansen APD, Hoogendijk E, Tulder M, Horst HE, Ostelo R. Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. Value in Health. 2015; 18(1): 35-43.
225. Rand S, Malley J, Towers AM, Netten A, Forder J. Validity and test-retest reliability of the self-completion adult social care outcomes toolkit (ASCOT-SCT4) with adults with long-term physical, sensory and mental health conditions in England. Health and Quality of Life Outcomes. 2017; 15(1): 163.
226. Kaambwa B, Gill L, McCaffrey N, LAncsar E, Cameron ID, Crotty M, Gray L, Ratcliffe J. An empirical comparison of the OPQoL-Brief, EQ-5D-3L and ASCOT in a community-dwelling population of older people. Health and Quality of Life Outcomes. 2015; 13: 164.
227. van Leeuwen KM, Jansen APD, Muntinga M, Bosmans J, Westerman M, van Tulder MW, van der Horst HE. Exploration of the content validity and feasibility of the EQ-5D-3L, ICECAP-O and ASCOT in older adults. BMC Health Services Research. 2015; 15: 201.
228. Jonker MF, Donkers B, de Bekker-Grob E, Stolk E. Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments. Health Economics. 2019; 28(3): 350-63.
229. Cook RD, Nachtsheim CJ. A Comparison of Algorithms for Constructing Exact D-Optimal Designs. Technometrics. 1980; 22(3): 315-24.
230. Fedorov VV. Theory of optimal experiments. Academic Press, 1972.
231. Walker JL, Wang Y, Thorlauge M, Ben-Akiva M. D-efficient or deficient? A robustness analysis of stated choice experimental designs. Theory and Decision. 2018; 84(2): 215-38.
232. Yang Z, Feng Z, Busschbach J, Stolk E, Luo N. 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. 2019; 22(7): 829-836.
233. Marten O, Mulhern B, Bansback N, Tsuchiya A. Implausible states: prevalence of EQ-5D-5L states in the general population and its effect on health state valuation. Medical Decision Making, in press.
234. Greene WH, Hensher D. A Latent Class Model for Discrete Choice Analysis: Contrasts with Mixed Logit. University of Sydney: Working Paper, 2002.
235. Pacifico D, Yoo HI. Lclogit: A Stata Command for Fitting Latent Class Conditional Logit Models via the Expectation-Maximisation Algorithm. The Stata Journal. 2013; 13(3): 625-39.
236. Hole AR. Fitting Mixed Logit Models by Using Maximum Simulated Likelihood. The Stata Journal. 2007; 7(3): 388-401.
237. Bhat CR. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B: Methodological. 2001; 35(7): 677-93.
238. Train K. Halton Sequences for Mixed Logit. University of California Berkeley, 2000.
239. Czajkowski M, Budziński W. Simulation error in maximum likelihood estimation of discrete choice models. Journal of Choice Modelling. 2019; 31: 73-85.
240. Train KE. Discrete Choice Methods with Simulation 2nd Edition. Cambridge: Cambridge University Press, 2005.
241. Gu Y, Hole AR, Knox S. Fitting the Generalised Multinomial Logit Model in Stata.. The Stata Journal. 2013; 13(2): 382-97.

242. Australian Bureau of Statistics. 2016 Census Quick Stats. 2017; Available from: http://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/036.
243. Stevens K, Brazier J, Rowen D. Estimating an exchange rate between the EQ-5D-3L and ASCOT. European Journal of Health Economics. 2018; 19(5): 653-61.
244. Flynn TN, Huynh E, Peters TJ, Al-Janabi H, Clemens S, Moody A, Coast J. Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. Health Economics. 2015; 24(3): 258-69.
245. Burgess L, Knox SA, Street DJ, Norman R. Comparing Designs Constructed With and Without Priors for Choice Experiments: A Case Study. Journal of Statistical Theory and Practice. 2015; 9(2): 330-60.
246. Burgess L, Street DJ, Wasi N. Comparing Designs for Choice Experiments: A Case Study. Journal of Statistical Theory and Practice. 2011; 5(1): 25-46.
247. Domínguez-Torreiro M. Alternative experimental design paradigms in choice experiments and their effects on consumer demand estimates for beef from endangered local cattle breeds: An empirical test. Food Quality and Preference. 2014; 35: 15-23.
248. Olsen S, Meyerhoff J. Will the alphabet soup of design criteria affect discrete choice experiment results? European Review of Agricultural Economics. 2017; 44(2): 309 - 36.
249. Street DJ, Mulhern B, Norman R, Viney R. Using simulations to compare DCE designs that could be used to value EQ-5D, in EuroQol Plenary, Barcelona, 2017.
250. Hole AR. Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment. Journal of Health Economics. 2008; 27(4): 1078-94.
251. Zwerina K. A general method for constructing efficient choice designs. Durham, NC, 1996.
252. Huang JC, Zhao MQ. Model selection and misspecification in discrete choice welfare analysis. Applied Economics. 2015;47(39): 4153-67.
253. Vermeulen B, Goos P, Scarpa R, Vandebroek M. Bayesian Conjoint Choice Designs for Measuring Willingness to Pay. Environmental and Resource Economics. 2011; 48(1): 129-49.
254. Burton, A., et al., The design of simulation studies in medical statistics. Statistics in Medicine, 2006. 25(24): p. 4279-4292.
255. Wang D, Li P. Does uniform design really work in stated choice modelling? A simulation study. Transportmetrica. 2005; 1(3): 209-21.
256. Johnson FR, Yang JC, Reed SD. The Internal Validity of Discrete Choice Experiment Data: A Testing Tool for Quantitative Assessments. Value in Health. 2019; 22(2): 157-60.
257. Kessels R, Jones B, Goos P, Vandebroek M. An Efficient Algorithm for Constructing Bayesian Optimal Choice Designs. Journal of Business and Economic Statistics. 2009; 27(2): 279-91.
258. Finch AP, Brazier JE, Mukuria C, Bjorner JB. An Exploratory Study on Using Principal-Component Analysis and Confirmatory Factor Analysis to Identify Bolt-On Dimensions: The EQ-5D Case Study. Value in Health. 2017; 20(10): 1362-75.
259. Mukuria C, Peasgood T, Brazier JE. Extending the QALY: Results of face validity and psychometric testing, in EuroQol Plenary, Brussels, 2019.