Valuing Wellbeing alongside Health with the Discrete Choice Experiment Method

Haode Wang

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

The University of Sheffield
Faculty of Medicine, Dentistry and Health
Sheffield Centre for Health and Related Research

July 2024

ABSTRACT

Quality-Adjusted Life Years (QALYs) encapsulate both the quality-of-life (QoL) and life-year gains from health treatments and cares. Preference-based measures are used to define health states that are then used to generate QALYs. Compared with the EQ-5D, EQ Health and Wellbeing instrument (EQ-HWB) and EQ Health and Wellbeing Short version (EQ-HWB-S) assess a range of QoL changes, including the impact on the health and wellbeing of care recipients and caregivers. Currently, there is no valuation study generating value set for EQ-HWB on the QALY scale. Discrete choice experiment (DCE) has grown in popularity in health state valuation, especially with long measures (attributes > 13). This study tested the feasibility of valuing the long health and wellbeing measure using the DCE method.

This study conducted literature review and qualitative consultations to find the most proper DCE design and information presentation strategy. The DCE with duration (DCE_{TTO}) design and DCE with a triplet dead state (DCE_{Death}) design were tested, with varied attribute order for the DCE_{TTO} to explore the order effect. A generator design approach applied for the choice set selection after simulation. 4056 UK and Australian general public completed the DCE survey. Demographic factors, overall QoL, health state preference, decision strategy and feedback were collected. Analysis using homogeneity and heterogeneity logit model, as well as constrained utility function, was conducted. The DCE_{TTO} data was anchored by calculating the marginal substitution rate to duration and the DCE_{Death} anchored by relative distance to latent dead state value.

All of the designs collected high-quality responses, where the proportion of negative feedback, strategic bias and inconsistency for the repeated question remained lower than 5%. The data analysis proved the feasibility of both DCE designs to generate QALY-scale value set. With the conditional logit regression result, this study found a varied order effect in UK and Australia. The DCE-Death had a larger proportion of health states defined as worse-than-death in both countries. A regression with constrained utility function generated value set with no insignificance and non-monotonicity, but utility distribution was systematically different from the individual effect models.

This study justified the feasibility of DCE in health and wellbeing preference valuation with empirical evidence. However, the design effect, model appropriateness and preference heterogeneity should be explored more in the future.

ACKNOWLEDGEMENT

I would like to thank my supervisors, Prof Donna Rowen, Dr. Clara Mukuria, Prof. John Brazier, Prof. Deborah Street (University of Technology Sydney) and Prof. Richard Norman (Curtin University), for their invaluable support and guidance throughout the design, development, and implementation of this research project. It is their encouragement, support and supervision that enabled me to muddle through the PhD journey, despite the challenges posed by the COVID-19 pandemic, on time. I am honored to have the opportunity to work with and be supervised by such esteemed mentors. They not only taught me how to conduct preference research, but their dedication and kindness have inspired me to keep the exploration.

I am also grateful to the individuals who contributed to my research through their discussions and insights: Dr. Anju Keetharuth and Prof. Rosalie Viney, who assisted with the ethics applications in the UK and Australia; Prof. Nick Bansback, who provided valuable discussion about the anchoring theory of ordinal data; Prof. Benjamin M. Craig, who reminded me to consider decision consistency; Prof. Zhihao Yang, who suggested testing other utility models; and Prof. Shunping Li and Dr. Paul Schneider, who offered valuable perspectives on statistical distribution. Additionally, I appreciate the overall comments on study design and final results from Prof. Feng Xie, Elly Stolk, Phil Powel, Brendon Mulhern, and Yuanyuan Gu. Their contributions were critical to the success of this PhD work.

Finally, I would like to express my gratitude to my family and my friends. Their support and understanding have been a precious source of strength throughout my PhD journey. It is always a pleasure to spend time with them.

This research was supported by the EuroQol Group. Thanks for their support and networking opportunities provided.

CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AIC	Akaike's information criterion
ABS	Australian Bureau of Statistics
AIHW	Australian Institute of Health and Welfare
ANOVA	Analysis of variance
BIBD	Balanced Incomplete Block Design
BIC	Bayesian information criteria
BWDCE	Best worst DCE
BWS	Best–Worst Scaling
BFFD	Blocked fractional factorial design
cTTO	Composite TTO
CFA	Confirmatory factor analysis
CMA	Cost Minimization Analysis
CBA	Cost-Benefit Analysis
CEA	Cost-Effectiveness Analysis
CUA	Cost-Utility Analysis
CALE	Cross-attribute level effect
DCETTO	DCE with a duration attribute
DCE-Death	DCE with a third dead state
DALY	Disability-Adjusted Life Year
DCE	Discrete Choice Experiment
EQ-HWB	EQ Health and Wellbeing
EQ-HWB-S	EQ Health and Wellbeing Short version
EORTC	European Organization for Research and Treatment of Cancer
EQ-VT	EuroQol Valuation Technology
EFA	Exploratory factor analysis
E-QALY	Extending the QALY
FIFO	Fold-in Fold-out
GPBM	Generic Preference-Based Measure
HWB	Health and Wellbeing
HRQoL	Health Related Quality-of-Life
HTA	Health Technology Assessment
HUI	Health Utilities Index
HYEs	Healthy Years Equivalents
HII	Hierarchical Information Integration
i.i.d	Identically distributed random variables
ICC	Intraclass correlation coefficient

ICER US	Institute for Clinical and Economic Review
ISPOR	The International Society for Pharmacoeconomics and Outcomes Research
MAEUT	Multi-attribute expected utility theory
MAUT	Multi-attribute utility theory
MNL	Multinomial logit
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
OWB	Objective Wellbeing
OLS	Ordinary least squares
PPIE	Patient and public involvement and engagement
PBAC	Pharmaceutical Benefits Advisory Committee
PAPRIKA	Potentially All Pairwise RanKings of all possible Alternatives
PBMs	Preference-Based Measures
QALYs	Quality Adjusted Life Years
QC	Quality Control
QoL	Quality of Life
RUM	Random utility theory
RUT	Random utility theory
RICHER	Rank Inclusion in Criterion Hierarchies with Extended Rankings
ScHARR	Sheffield centre of Health and Related Research
SF-6D	Short-form 6-dimension
S.D.	Standard errors
SG	Standard Gamble
SWB	Subjective Wellbeing
TTO	Time-Trade-Off
ONS	UK Office for National Statistics
UNDP	United Nations Development Program
VAS	Visual Analogue Scale
vNM	Von Neumann–Morgenstern
WEMWBS	Warwick-Edinburgh Mental Wellbeing Scale
WELBY	Wellbeing adjusted life years
WTP	Willingness-To-Pay
WTD	Worse-than-dead
ZBT model	Zermelo Bradley Terry model

RESEARCH OUTPUT

The following outputs are based on the work presented in this thesis. None of the thesis chapters are presented in the publication format, as not all of the contents presented here are included in the published version.

Publication:

Wang, H., Rowen, D. L., Brazier, J. E., & Jiang, L. (2023). Discrete choice experiments in health state valuation: a systematic review of progress and new trends. *Applied Health Economics and Health Policy*, *21*(3), 405-418.

Conference presentation, seminars, and talks:

Haode W. Discrete Choice Experiments in Health State Valuation: Progress and New Trends. [Presentation] June 2023. 7th National Patient Reported Outcome Measures (PROMs) Research Conference. Sheffield, UK

Haode W. The EQ-HWB Valuation: A methodology and feasibility study. [Presentation] June 2023. International Academy of Health Preference Research Meeting. Sydney, Australia

Haode W. The EQ-HWB Valuation: A methodology and feasibility study. [Workshop] July 2023. Centre for Health Economics Research and Evaluation, UTS group meeting. Sydney, Australia

Haode W. Valuing health and wellbeing using DCE method: a feasibility study with EQ-HWB measure. [Presentation] June 2024. ScHARR PGR Conference. Sheffield, UK

Haode W., Rowen, D. L., & Mukuria, C. Valuing a long measure of Health and Wellbeing from EQ-HWB using Discrete Choice Experiments (DCE): A feasibility study of two methods with UK and Australia general public samples. [Poster] September 2024. 41st EuroQol Plenary 2024. Noordwijk, the Netherlands.

THESIS OUTLINE

This study provided feasibility evidence of valuing health and wellbeing with the EQ-HWB measure by using EQ-HWB instrument for the first time. A systematic exploration was reported in the following 8 chapters, including literature review evidence, qualitative consultation results and an international large-scale survey in two English-speaking countries (UK and Australia). The feasibility is evaluated by the data quality, feedback and the regression results. The study also targets on three research questions around health and wellbeing DCE valuation study design: *does the attribute ordering of health and wellbeing affect choices and answering times, what is the preference effect of different DCE design*s and *the HWB preference differ for public samples from the UK and Australia.* By exploring the three questions, the results will benefit future valuation design by providing essential evidence on information presentation, DCE design choice and preference heterogeneity.

Chapter 1 and 2 provide a general overview and introduce relevant concepts including economic evaluation, the QALY and current methods used to generate QALYs, and the EQ-HWB measure. Chapter 3 reports a systematic review on the DCE literature on health state valuation and considers how these methods can be applied to value EQ-HWB. The literature review has found an increasing interest in this valuation method in the recent years and four DCE design strategies that might be feasible for the health and wellbeing valuation study.

Chapter 4 introduces the systematic selection of attributes from EQ-HWB to take forward to valuation. A systematic approach based on dimensionality, item performance, stakeholder preference, international and cultural feasibility, and qualitative evidence lead to the selection of 13 out of the 25 EQ-HWB attributes for valuation using DCE.

Chapter 5 undertakes qualitative research using focus groups to inform the study design. The focus groups assess the feasibility of valuing the large number of attributes selected for valuation and confirm task wording, presentation and design. This qualitative research confirms that the set of selected EQ-HWB attributes can be understood and valued in both paired DCE (DCE_{TTO} where participants choose which of two hypothetical EQ-HWB states they prefer where each has a specified duration before death) and triplet DCE (DCE-Death, involving two hypothetical EQ-HWB states and a triplet dead state, where participants choose which is best and which is worst) formats, though some presentation formats have not appeared suitable. The focus

group feedback is used to select and refine the DCE valuation design, retaining both paired and triplet designs with overlapped and color-coded attributes.

Chapter 6 details the development of the online DCE survey conducted in the UK and Australia. Members of the general public – a population suggested by the majority of HTA agencies - have been selected. Feasibility of eliciting preferences using the paired and triplet DCE tasks has been investigated, and three further research questions are investigated: 1) the influence of presenting information in different orders (health attributes first or wellbeing attributes first); 2) the task design effect on elicited preferences (compare DCE_{TTO} design and DCE-Death design); and 3) the UK and Australian general public health and wellbeing preference differences, by designs. The first research question is to examine the impact on preferences by comparing the results of the DCE_{TTO} presenting the attributes in an EQ-HWB order (health attributes first, then wellbeing attributes, followed by pain severity attribute) versus a revised-order prioritizing wellbeing attributes (wellbeing attributes first, then health attributes, followed by pain severity attribute). The second research question is to examine the impact from using the different task designs - DCE-Death design and DCE_{TTO} designs — which introduce alternative design and anchoring strategies, including anchoring with duration or with just dead dummy for generating the full health- death scale value set. A generator design has been used to select the choice sets for the DCE. The third research question is to examine the difference in modelled preferences with samples from UK and Australia for both tasks.

Finally, Chapter 7 presents the data analysis results. Both DCE designs are feasible, despite the logical consistency for wellbeing attributes being low. The EQ-HWB-order and revised-order have significant preference effects on utility decrements for Depression, Hearing, Getting around inside and outside, Control in the regression considering the order effect, but the magnitudes are close to zero. Significant variance is observed with the two DCE tasks (DCE_{TTO} and DCE-Death) and across the two countries. UK and Australian samples generate varied point estimation for utility values, but the general trend is similar. An overall discussion of the research findings, implications for future research and study limitations are provided in Chapter 8. The findings support the future valuation of the long EQ-HWB measure (the full 25 attribute version) using the DCE method, while highlighting the need to consider potential biases in methodology selection and data modelling.

Chapter 1 Health State Valuation: from What to How

Health state valuation provides essential quality-of-life information for decision modelling and decision making. Utility calculation requires a health (and/or wellbeing) quality-of-life (QoL) descriptive system and value set generated from the target population. To evaluate a broader health change, there has been a shift from using descriptive system merely considering physical and mental health, to broader health and wellbeing benefit. In the recent years, valuation studies using time trade-off and /or discrete choice experiment valuation design have shown high feasibility and validity[1], but more valuation study design options has been mentioned before. This chapter provided an overview of the utility theory: what is utility, what are the main instruments for describing it and the valuation methods.

1.1 Economic Evaluation of Healthcare

For all the stakeholders in the health sector who provide, regulate, pay, and receive the healthcare in either central-planning healthcare system or a free market, resource scarcity influences their decision making and accessibility to necessary services[2]. The health market struggles to distribute resources efficiently due to the abnormal features of health products and information asymmetry between providers and consumers[3, 4]. While Adam Smith's "invisible hand" metaphor highlights self-interest in a free market, health economic evaluation represents a more "visible hand," guiding the optimization of the healthcare market in an imperfect world[5]. Since 1999, the National Institute for Health and Care Excellence (NICE) has been assessing the cost-effectiveness evidence of new health technologies[6, 7]. The economic evaluation frameworks are Cost-Effectiveness Analysis (CEA), Cost-Benefit Analysis (CBA) and Cost-Utility Analysis (CUA) [8, 9]. CBA, used in both public policy and business, quantifies the costs and benefits of various interventions in monetary terms. CEA compares the costs and health outcomes of different health interventions, though healthcare providers may use varied outcome measures or indicators. CUA, a subset of CEA, uses Quality-Adjusted Life Years (QALY) as a universal measure of benefit, which is an arithmetic product of length and quality of life change[10]. In other word, ‘Benefit’ in CUA steps forward to a two-dimensional unit capturing both quality and quantity of life[11]. This approach allows for theoretical cross-disease comparisons (Table 1). A key feature of decisions made using CUA is to focus on maximizing overall utility through efficient allocation of healthcare resources. The definition and calculation approach of QALY shape the vision and mission of healthcare systems.

Table 1 Comparison of Economic Evaluation Methods

Method	Description	Cost measure	Effectiveness measure	Measure outcome
CEA	Consider the costs and common effects difference of alternative interventions	Monetary term	Clinical units (e.g., progression-free survival, life-year gain)	Incremental cost per effectiveness improvement
CBA	Translate effects into the monetary unit and consider with intervention costs	Monetary term	Monetary term	Net monetary benefit
CUA	Using the generic measure (e.g., QALY) to express health gain	Monetary term	QALY, DALY	Incremental cost per QALY gain
CMA	Compare the cost of two or more alternative interventions, under consideration that outcome is the same	Monetary term	None	Health resource saved in monetary term

1.2 Measure of Health

1.2.1 Measure of health gain

Before introducing QALY and its evaluation, it is helpful to begin by considering all the health gain measurement concepts and why QALY is preferred. Healthy Years Equivalents (HYEs)[12] accounts for the gains in mortality (through extended life expectancy) and quality of life. It quantifies the number of years in perfect health that are equivalent in utility to the health states in the fixed sequence under consideration[4]. HYE incorporates preferences related to sequence and time. An alternative to QALY and HYE is the Disability-Adjusted Life Year (DALY), developed by the World Health Organization (WHO)[13]. DALY quantifies health loss using disability weights derived from expert panels or general population surveys[13][14]. However, DALY focuses more on quantifying the burden of diseases of the society, instead of catering to measure broad social welfare changes.

QALY measures '*health outcome which assigns to each period of time a weight, ranging from 0 to 1, corresponding to the health-related quality of life during that period, where a weight of 1 corresponds to optimal health, and a weight of 0 corresponds to health state judged to be equivalent to being dead*[15].' The health economists defined the utility values of each corresponding health state as the numerical value of the QoL. Namely, the researcher uses a cardinal value to represent the general life wellness (or health wellness for HRQoL). QALY relaxes HYEs' assumption by calculating the health of certain period independently[4, 16]. The total QALYs for an individual is calculated by multiplying the health state values with the duration spent in each state. The health

state values, expressed as explicit “utility” to represent the invisible satisfaction of each health state{Fallowfield, 2009 #939}. The health states are normally described by a number of attributes, while a utility value is needed.

There are critiques that the QALY represents utility based on unrealistic assumptions. For instance, general public evaluates health state utility in a discontinuous way and order of health states influence their satisfaction with the general health[17]. However, in practice, using QALY to calculate the overall change of certain technology is more feasible than using HYE or DALY, as HYE's reliance on the sequence of health state presentations challenging the validity of an universally health benefit for same patient group[18] and DALY is less flexible in describing health states. Measuring health gains with QALY in economic evaluations is considered the most practical approach, even though it's not perfect.

1.2.2 General public preference V.S. patient preference

QALYs encapsulate both the health-related quality-of-life (HRQoL) and life-year gains from health treatments. The life length is straightforward for measurement, while the HRQoL of each health state can never be observed directly[19]. A utility value that represents HRQoL on a 0 (denoting death) to 1 scale (denoting perfect health) normally used. The most commonly used approach for estimating utility values is to elicit the preference for each health state from certain population. Since it is impractical to calculate the arithmetic mean of every individual, a crucial question arises: who should be the representative sample for calculating the preference-based utility values in QALY calculation?

Empirical evidence shows a systematic difference between utilities derived from patients and the general public. One review concludes that patients often assign higher utility values, possibly due to their familiarity to specific health conditions and consideration of adaptation, while the general public focus more on the negative aspects[20, 21]. This divergence leads to a lack of consensus on which group's preferences should guide health resource allocation. Utilizing public preferences has its merits, such as representing the interests of taxpayers and potential voters, who are the 'real payers' of publicly funded healthcare, and the ‘veil of ignorance’ that the healthy public representatives have complete impartiality of judgment[15][4]. Compared with eliciting preference of patient groups by separate surveys, valuation with one general population would be more practical[22]. However, a drawback is that healthy individuals may lack an understanding of the physical and emotional impacts of illness and the process of adapting to chronic diseases[23].

The choice of whose preferences to use is a normative issue for national HTA agencies. In countries such as the UK, Canada, and the Netherlands, where a significant portion of health funding comes from taxes, HTA agencies often recommend using the general public's preferences[24]. In contrast, the Swedish Dental and Pharmaceutical Benefits Agency favors patient perspectives in cost-effectiveness analysis[25]. The Washington Panel on Cost-Effectiveness in Health and Medicine in the U.S., suggests treating patient and public preferences equally[15, 26]. Some researchers, like McTaggart *et al.,* and Clarke *et al.*, proposed middle-ground approaches, such as using patient audio descriptions to inform the general public during surveys, or providing detailed information about specific diseases before utility evaluations[27, 28]. However, these methods have not gained widespread acceptance. Despite ongoing debates about the ideal target population, this research conducted the survey from a general public preference perspective, to maintain a population consistency with the HTA institute recommendation in UK and Australia.

1.2.3 “Health” in QALY

A description of health states or descriptive system is needed for evaluating the QoL[29]. Early method is to construct bespoke descriptions of 'imperfect' health states (or vignettes). These vignettes are usually developed through interviews with clinicians, patients, or the general public, often targeting less common scenarios like rare diseases[30]. However, HTA institutes less favor vignette or self-health descriptions due to comparability issues.

A more widely accepted method is using preference-based measures (PBMs) of HRQoL, where each state is described by a fixed number of attributes introduced by the measure. A PBM consists of a patient questionnaire, a health state classification system, and the value set[4]. The utility score for each health state can be obtained through preference weights. The subjective impact of health states on the quality of life and societal desirability for each health state can be quantified by PBMs. EQ-5D (the 3 level version EQ-5D-3L and the 5 level version EQ-5D-5L) is NICE’s current reference instrument. However, due to the limited number of attributes, the 'health' definition (and subsequently the QALYs generated) is narrow{Torrance, 1987 #842}. EQ-5D has been criticized for lacking sensitivity to certain conditions or illnesses, such as dementia, hearing, vision, and broader wellbeing impacts[31]. Short-form 6-dimension (SF-6D) and Health Utilities Index (HUI) are similar health measures. There are increasing concerns that current generic PBMs may be not appropriate for mental health[32], social care[33], and public health interventions[34].

To measure the QoL change beyond traditional health notion, measures such as the capability wellbeing measure ICECAP-A[35], social-care quality of life measure ASCOT[36], and Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS)[4] have been developed, focusing on broader aspects of wellbeing. These are intended to be used alongside traditional health measures in HTA contexts. However, combining health QALYs and wellbeing outcomes (wellbeing adjusted life years, WELBYs) into a single score is problematic. Using multiple measures leads to challenges in synthesizing analysis and making cross-program comparisons among healthcare interventions, social care, and public health.

To encompass a more inclusive evaluation of health gains, a compromised approach is to use adjusted QALYs[4, 37], where a societal preference weight is attached to different kinds of QALY gains. This adjustment, however, requires extensive societal data and challenge the “A QALY is a QALY is a QALY” assumption[38]. The most obvious limitation is the equality concerns. A more ambitious but potentially more effective way is to evaluate broader health and wellbeing change with one measure. The concept of 'Q' in QALYs beyond just HRQoL or WELBYs[39]. The target measure (EQ-HWB) used in this exploratory study evaluates both health and wellbeing through single measure. The QALYs generated is an “extended” QALY evaluating both health and wellbeing, compared to other wellbeing measures and HRQoL preference-based measures. A more detailed introduction to the EQ-HWB measure can be found in Chapter 2.

1.2.4 Definition of Wellbeing

There are many different theories of wellbeing to choose from. Sen conceptualizes wellbeing as the ‘beings and doings[53]’, moving from the economic growth to the ability of individuals[54]. His framework of wellbeing concentrates on the human capabilities and freedom people have to achieve valuable functioning on practice [55]. This framework exhibits significant relationships between capacities of physical health (what people can do) and the all-inclusive concept of QoL[40]. A wellbeing improvement indicates human development of expanding capabilities in multiple aspects, for example, physical health ability improvement, better mental health condition and support by other people to achieve one thing [56].

On the other hand, Ryff conceptualized psychological wellbeing by the primary positive facets of life: environmental mastery (control of environment), positive relationships, autonomy, personal growth, self-acceptance and purpose in life (self-expectations), implies that optimal functioning and life satisfaction are metrics of value in assessing

wellbeing, which is beyond the subjective wellness (hedonic well-being, life satisfaction, and eudaimonia) but more specific than Sen's definition [57, 58, 59]. Ryff has developed

Subjective Wellbeing (SWB) is a broader definition than hedonism with various measures considering life satisfaction, satisfaction with life aspects (feeling of happiness and sad and pain), eudemonic wellbeing (sense of purpose and meaning) beyond health could be defined as wellbeing [61]. By comparing with Ryff's wellbeing theory, the main difference is that SWB admits a significant difference on wellbeing definition between culture groups and communities and does not rely on an universal judgement or measurement scale[62]. SWB admits the influence of personal expectations and frames of reference on general wellbeing [48, 50].

To evaluate the wellbeing from a more objective perspective, the United Nations Development Program (UNDP) defines an Objective Wellbeing (OWB) by six major dimensions: health, job opportunities, socioeconomic development, environment, safety, and politics, where OWB can be assessed through the extent these needs are satisfied (observed needs satisfaction)[63]. The definition of OWB bears a resemblance to Sen's concept of the "resources needed" for wellbeing.

In conclusion, the definition of wellbeing varied but there are some similarities: the wellbeing is beyond immediate pleasure or subjective feelings and should be considered with health and capabilities necessary for individuals to achieve their desired 'beings and doings{Sirgy, 2012 #941}.' As summarized by, "there is a plethora of concepts directly related to (subjective) wellbeing" in the field of quality of life and it significantly influence people's wellness. For the QoL evaluation, wellbeing is not simply a psychological matter but an evaluative matter that concerns the comprehensive evaluation of utility change.

1.3 Health State Valuation

As described above, our discussion is about the value of "health" that is described by a PBM. Brazier *et al.* describe the process of health state valuation using PBMs as "*assigning disutility weights to each attribute level*"[4]. In other word, the health state valuation generates preference weights or QALY weights[41] on each of the health or wellbeing dimensions, representing a relative preference on each dimension and dimension levels. This section provides a narrative review of the valuation methods.

To describe health aspects, the terminologies employed by research are dimensions, attributes, aspects and factors. Although they are not equated sometimes, the

demarcation may not be clear in the original research. In this section, all of the terminologies share the same meaning of aspects (or certain aspect of health). Similar with health aspect terminologies, profiles, health states and scenario share the same meaning of a description of certain health state in the valuation tasks.

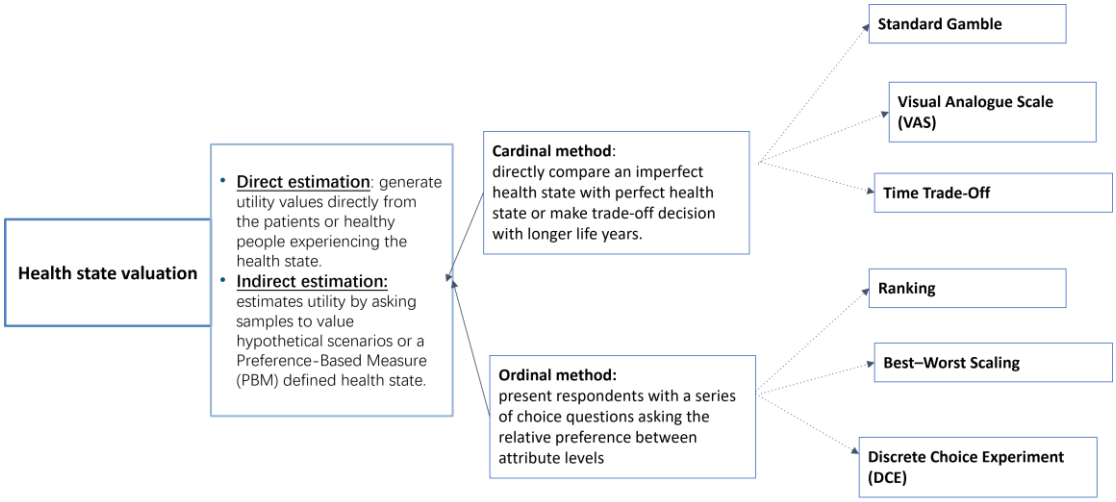
1.3.1 The definition of health state valuation

The health state utility values are scaled from 0 to 1, where 1 represents optimal health and 0 equates to death or the worst possible health state[42]. A negative value represents the health state is generally considered worse-than-dead (WTD). With a given descriptive system PBM, the value set maps any health state onto a utility value – also referred to as scores or weights – by eliciting preferences for different health states from general public participants[43].

In the context of economics evaluation and HTA, the most commonly used methods are cardinal methods Time-Trade-Off (TTO), Visual Analogue Scale (VAS), and Standard Gamble (SG), or ordinal methods ranking, Best–Worst Scaling (BWS) and Discrete Choice Experiment (DCE).

Cardinal methods ask respondents to directly compare an imperfect health state with perfect health state or make a trade-off decision with longer life years. The ordinal methods present respondents with a series of choice questions with systematically varied attribute levels. While there's no consensus on the best valuation method, cardinal methods are often recommended by valuation protocols. However, ordinal methods have gained more attention recently as their choice tasks are perceived to be easier for respondents to understand than classic cardinal methods[4, 44, 45]. The following sections will delve into the cardinal method and then explore the main ordinal methods (Figure 1).

Figure 1 Health state valuation method



1

1.3.2 Standard Gamble

The SG is a classic preference elicitation method for direct and indirect health utility elicitation derived from the von Neumann–Morgenstern (vNM) utility theory with the first-order utility independence assumption[42], which presupposes decision-making aimed at maximizing utility under the axioms of preference transitivity, independence, and continuity[46, 47]. Respondents are asked to compare a certain health state (State A) with a probabilistic combination of two uncertain health states (State B and State C). State B (normally perfect health) is better than State A and State C (normally worst health state or death) is worse than State A. In practice, the B and C states are full heath and death to make anchoring onto the zero to one scale required to generate QALYs easier. The probability of perfect health (p) and death ($1-p$) are varied until the respondent confirms indifference for the two options[4, 42]. The indifference probability of perfect health (p) is the QALY scale cardinal value for chronic health states better than death. If the health state h_l is expected to be worse than death, then the task is altered such that State C would be the valued health state and full health remained the same, while the State A should be death[48].

SG is a classic and reliable method with high practicality that the completion rate of SG was expected to be over 80%[4]. However, regarding SG as a gold standard has been criticized. The gambling and probability aspects could be cognitively complex for certain respondent groups[49], as the probability trade-off is relatively abstract. Respondents cannot fully understand the meaning of ‘uncertainty’ in the gambling procedure. SG assumes respondents to be risk-neutral, where the respondent’s risk attitude may be a mixture of risk-averse, risk-neutral, and risk-seeking [4, 50].

1.3.3 Time Trade-Off

TTO tasks elicit the health state utility with an iterative procedure that reflects the 'sacrifice' idea in economic reasoning[51]. Compared with SG, TTO incorporates no health state probability in the preference elicitation task. It asks respondents to trade-off between two certain health states with iteratively changing duration, rather than gambling the lotteries with various probabilities. For health states better than death, the two alternatives would be living in an imperfect health state for t years (e.g., t equals ten years) and living in full health for x years (x<t). X is changed iteratively until the respondent achieves indifference between the two alternatives[42]. Under the MVH protocol, the York Measurement and Valuation of Health (York MVH) protocol[52] for health states worse than death, the first alternative would be a ten-year lifetime made up of living in an imperfect health state for x years, followed by living in full health for the rest of the lifetime. The other alternative is immediate death, which means the utility is zero. In other word, this question asks respondents to find the balanced point for achieving 0 utility with a health state worse than death and perfect health. The health state value in the QALY scale is x/10 for states better than death and 1-(10/x) for states worse than death[4]. Researchers have developed optimized TTO study design strategies called lead-time (respondents imagine starting with a period of full health before experiencing the imperfect health state), or lag-time (an imperfect health state followed by a period of full health before immediate death) TTO design[53], or a combination of traditional TTO and lead-time TTO[54].

One risk with TTO is the time factor in expected utility theory[51], where the valuation results tend to be confounded by individual time preference. The QALY gains are not linearly increasing with the time spent in the state[55], and the poorer health states become more intolerable with a longer fixed duration[56]. If changing the decision background information from 10-years to longer or shorter life expectancy, then the result for worse health states will be changed. Besides, respondents still have a strong cognitive burden on distinguishing health states if the attribute levels are similar, and have varied interpretation on the painful feeling of 'death'[57].

1.3.4 Visual Analogue Scale (VAS)

VAS is a preference measurement technique applied in market research, social science, and health care[42, 58, 59]. There are many variants of the VAS technique in line format and question-wording. One example is where respondents are asked to mark the corresponding VAS scale for the health state on a line with the best and worst state as the benchmarks on the two ends[4]. Each point on the line represents a

correspondent score of value of the health state. By using VAS values on a 0 (death) to 1 (full health) QALY scale, respondents need to place the value of death as well. The best health is anchored as 100 and the worst (or death) as 0, with 99 marks between the two ends. The single number on the VAS scale represents relative QoL value. The latent values can be anchored to the 0 to 1 scale by dividing with 100.

VAS has no trade-offs or gambling process to compare one or more alternatives in valuation, and has been used to value the Quality of Wellbeing Scale (QWB) and EQ-5D measures [4, 60]. However, there are concerns about using VAS to value health states due to different biases. Similar health states could be placed on distant positions [61], and the VAS values are more accurately reflect the ordinal relation of health states instead of cardinal latent QoL values: participants may not perceive equal differences though the assumption behind is each unit change on the scale representing the same QoL change[53, 62]. Besides, respondents tend to avoid putting the health states near 100/0 of the VAS scale, such as selecting numbers ending in 0 or 5 (e.g. 50, 35), namely end-of-scale bias. Considering other comparative methods, VAS involves no trade-off process in generating the “preference utilities”, where some studies challenged whether the utility values were the same[63].

1.4 Using Ordinal Data to Value Health State

This section introduces the main ordinal methods including DCE, which is the focus of this thesis. There are many ordinal data types, including ranking, rating, DCE, and BWS. Ordinal health state valuation methods are based on the assumption that attributes can describe any objective or option, and the utility can be calculated with a finite formulation of attributes [42, 64, 65]. The theoretical foundation of ordinal data analysis is random utility theory (RUT), which posits that selection is a stochastic process with rational evaluation and random errors[66]. The consumer’s (patient’s for the healthcare preference elicitation studies) utility is derived from a limited number of consistent characteristics, instead of presenting inconsistent description of each option[67].

1.4.1 Ranking

Ranking asks respondents to provide a complete order of the possible hypothetical heath states in health state valuations[4]. Respondents need to select the best health state out of a series of states, followed by identifying the second-best state and continue until all the states are ranked. Dolan *et al.* models rank health state valuation data using the conditional logit model[68].

One characteristic of ranking is the low cognitive burden and consistency with other choice methods{Marra, 2007 #940}. Ranking has long been regarded as having low cognitive burden and is sometimes used as exercise questions or warm-up exercises, in order to familiarize respondents with the hypothetical health states[69]. Early studies concluded that the valuation results had no significant difference between ranking and single choice questions with those options ranked high[70]. However, as the number of attributes in each health state increases, ranking can be hard to apply. A second disadvantage is accuracy. If health states variations are small, or the respondents cannot distinguish the attribute levels from each other, the random error term becomes the decisive factor[71].

1.4.2 Best–Worst Scaling

BWS is a priority assessing method to determine the best and worst among a set of items[72]. There are three types of BWS: case I BWS (object case) presents a series of attribute levels by asking people to select both the best (most important) and worst (least important) options[73]. The case II (profile case) presents a single description of health state. The respondents are asked to select the best and worst piece of information with the given description. Case III is an extension of a discrete choice, by presenting three or more profiles to choose the best and worst profiles[74].

BWS applications include the valuation of the wellbeing measure ICECAP[75] and the child health measure CHU-9D[76]. Osman *et al*. showed that BWS generated stable preference with better test-retest reliability than other ordinal methods[77]. However, its consistency with other ordinal methods, such as DCE and ranking, seems to be lower. A noteworthy rebuttal to this common consistency criticism was the argument that other choice-based methods were not superior on the societal level[78].

1.4.3 Discrete Choice Experiment

DCE and its data modelling method is proposed by Louviere *et al* and defined as a “general preference elicitation” survey approach[79, 80]. The method obtained small amount of information from each participant to estimate a general social preference[81]. DCE asks respondents to make a choice (e.g., *which health state do you prefer*) between two or more alternatives where at least one attribute is ‘systematically varied’. It assumes that the respondents have a latent utility function which maps the characteristics (attributes) of a given option to a certain value.

PBMs sources the attributes for scenario description. Hakim and Pathak reported the first DCE study measuring health state preferences[82] in 1999. Another example is to

estimate the disutility weights of the child Behavioral Problems Index (BPI), where researchers asked respondents to compare two scenarios for a 7-year-old child[83]. Using DCE method for health state valuation attracted attention in the recent years, as it is considered easy to apply. A 2018 literature review identified 63 published studies using the DCE method, and there was a steady increasing trend[84]. EuroQol group EQ-VT protocol (version I and version II) recommended both TTO and DCE in valuation studies[53]. However, there is no gold standard in DCE design[85], data analysis[86], anchoring of latent utility values[87] and its comparability with the cardinal valuation.

1.4.4 Design an ordinal preference elicitation study

The three methods discussed above are all choice-based ordinal method (i.e. respondents need to make a choice). An important point to note is that all methods yield different values but share similarities in the valuation study design. Given the significance of the valuation study design of ordinal method, a better understanding of the necessary steps, and clearer explanation of its difference with cardinal method, would be helpful.

First, a descriptive system is devised or selected that defines the evaluated health states[33, 88]. The attributes and levels are from existing measures, expert interview, stakeholder interview, pilot test, and self-creation[75, 89]. There is no strict standard for the maximum number of attributes in valuation study. However, a larger amount of information increases study design complexity and respondent's cognitive burden[90].

Secondly, the choice questions are constructed. A full factorial method is to present all of the possible pairs, or fractional factorial designs for evaluation[85]. Since the number of scenarios may exceed the maximum number[91], researchers developed fractional factorial design to "select" a subset of scenarios for survey design. The selection criterion are statistical efficiency (the efficient design, or named optimal design)[92, 93], orthogonality (orthogonal design or generator design)[94] and random selection. Several constructions of efficiency evaluation standard have been developed and tested by statistical studies, including minimizing the generalized variance of all dummy-coded factors (D-efficient design), minimizing the arithmetic mean variance of the inverse of the information matrix (A-efficient design), and minimizing the variance of best linear unbiased estimator (C-efficient design)[95-97] (Table 1, Appendix A). The most commonly used method was D-efficient design. However, a test should be evaluated from the perspective of the design explanation power and the efficiency with given sample size[98, 99].

Thirdly, the presentation strategy is decided. The practical consideration justifies the practical aspects, such as the number of questions per respondent and the presentation format. DCE recommends an optimal question number per respondent as less than 15[100] for intervention preference study, and 12 for health state valuation studies[84]. Arguments are made for the use of introduction/warm-up questions[101], interactive information presentation, qualitative consultation and pre-test/pilots[102].

The last step is survey implementation. One should make the subjective justification that whose assessment informs the research question, and derive a value set from the representatives[11]. Note that there is no particular concept provided to underlie the sample size. An optimal design should have at least 20 respondents per version of the choice set [103]. A statistically reliable method is to consider coefficient confidence level, statistical power, coefficient covariance, effect size[104] and data collection mode[105], where the collection mode, such as online data collection, affects the data quality and missing rate.

1.4.5 Ordinal data modelling

Analysis of discrete choice data refers to the process of modelling quantitative dependent variables with linear or non-linear expression. The utility function should be decided by considering: (i) the experimental design and identified effects (main/interaction effects), and (ii) the number of profiles in each choice set[11]. In general, an additive function can calculate the observed utility, with marginal effects (level change) being weighted by a unique parameter. Model fitness can be informed by Log-likelihood and pseudo-R-squared values. Y in Formula 1 represents the latent utility value, which is a linear additive function of the attribute levels (X_{ij}) and the error term ϵ_i for individual i and option j for a certain choice set. More information about data modelling and the utility function can be found in the following session.

$$Y = \beta_i' X_{ij} + \epsilon_i \qquad \text{Formula (1)}$$

Preference-based measures are consisted of multiple attributes and the value sets are built on the multi-attribute utility theory (MAUT), as put forth by Lancaster[106, 107], posits that the utility of a certain product (e.g., intervention or healthcare service) can be determined by preference value function of objectives and decision uncertainty incorporated[108]. This makes it possible to predict the preference values for different alternatives. MAUT underlies the theory of conjoint analysis utility (also known as multi-attribute value), which assumes that preferences are determined under specific, riskless circumstances[71]. The difference between MAUT and multi-attribute

expected utility theory (MAEUT), normally used with standard gamble data analysis, is utility value of one MAEUT alternative is a function of a series of 'expected' health state utility, with an uncertainty probability attached [115]. Given the estimation difficulty, MAUT is always preferred.

In choice-based health state valuation research, the data analysis lies in this framework. The preference value function, alternatives, objectives and performances are utility function, health states, dimensions and levels[109]. The utility function may consider both quality and quantity of life. To give an example, a classic DCE utility function consists of health state profile and duration attribute[110]: t_{ij} is the expected time in given health state, X_{ij} is the dummy-coded vector of attribute levels describing the health state. $\beta_1$and $\beta_2$are the estimated latent utility associated to each factor. The data is modelled with a constant proportional time trade-off assumption. Utility for a given health state U_{ij} is a function of quantity of life $\beta_1 t_{ij}$ and the interaction effect of quality and quantity of life $\beta_2' X_{ij} * t_{ij}$, with an additional error term to reflect the random effect (Formula 2)[44].

$$\mathrm{U}_{ij} = \beta_1 t_{ij} + \beta_2' X_{ij} * t_{ij} + \varepsilon_{ij} \qquad \text{Formula (2)}$$

Two optimizations with the simplified model are accounting for the heterogeneity and heteroscedasticity of individual preference. The rational for homogeneous assumption is that ordinal data can capture individual heterogeneous preferences but should investigate the general preference from observed homogeneous population[71]. However, this assumption can be too restrictive for more advanced applications beyond population-level health economics research, such as the individual simulation[111]. The preference heterogeneity, where the individual-level preference variance correlated to their personal characteristics accounted, can be investigated by using a mixing distribution of parameters to represent the uncertainty around each individual[112-114], named mixed logit model, or classified the respondents to various classes by considering the individual characteristics that can be correlated to their attribute preference, named latent class model[115]. With the heterogeneity model, the inherent Independence from Irrelevant Alternatives (IIA) assumption holds. Heteroscedastic models relax the IIA assumption by using flexible error terms related to population groups. Namely, the uncontrollable and unobservable difference between options (e.g., option difference) and/or between options (e.g., order of choice sets) leads to dependent and differently distributed unobservable preference factors (the error terms)[116]. The heteroscedastic models, such as the nested logit model or probit models, are less commonly used as the variance-covariance correlations are hard to

explain. Thus, it has become practical and important to investigate whether the bias caused by homogeneity or IIA assumptions is significant by using the mentioned models. As the health state valuation targets on eliciting the preference of targeted attribute levels and the order for presenting health state in DCE is normally randomized, an exploration with preference heteroscedasticity is less common.

1.4.6 Normalization scaled values

The process of normalizing scaled values in health economics, particularly for QALY calculations, involves several key steps and methodologies to ensure that derived utility values from discrete choice data fit within the standard 0 to 1 scale. The normalization can be summarized as a positive affine transformation, where γ2 is the best health state utility value and γ1 is the "normalization" constant for the latent disutility calculated by $\beta X' + \varepsilon_{ij}$ (Formula 3)[117]. The γ_2 represents the utility value of the best health state with a particular descriptive system, and γ_1 is a rescaling constant for the latent disutility. The predicted utility for a given health state would be $\gamma_1 \beta X' + \gamma_2$. This thesis introduces five main methods that have been successfully applied before, namely, how to derive the γ_1.

$$U_{ij} = \gamma_1 (\beta X' + \varepsilon_{ij}) + \gamma_2 \quad \text{Formula (3)}$$

1) Anchoring with hybrid model: This model calculates level coefficients by maximizing the likelihood of observations from both DCE and TTO tasks while minimizing the possibility of logical inconsistency[118]. In this model, it is assumed that an individual's TTO and ordinal responses reflect identical preferences for health states. This allows for a linear transformation of the ordinal value function to the TTO value function using the transformation parameter θ, thus enabling a common parameter for each attribute level (Formula 4)[119].

$$\beta_{TTO} = \theta \times \beta_{DCE} \quad \text{Formula (4)}$$

2) Anchoring with the coefficient of 'dead': The second method uses a dummy variable for 'death' in regression models, rescaling attribute level coefficients by dividing by the coefficient of the 'dead' dummy. This method inspired by utility anchoring research with ranking data, where the utility value anchored onto 0 to 1 through rescaling with an additional death dummy coefficient[87, 120, 121], but not directly for generating a DCE value set. The function for applying this method with DCE data will be introduced in Chapter 6.

   A notable implementation of this method is with UK EQ-5D-3L data, where both

DCE and TTO data collected. In this hybrid model, the DCE pairs two health states, A and B, with an option to substitute state B with the immediate death state C, without specifying duration[87]. Within each DCE block, 90% of the questions compare health states A and B, while the remaining 10% introduce a comparison between a very poor health state and immediate death, allowing for the direct extraction of better-or-worse-than-death preferences alongside the comparative valuation of two health states within a single DCE framework.

3) Anchoring with cardinal value: the third method utilize exogeneous health utility values generated by the cardinal methods. The first option anchors scale's lowest end at the cardinal value of the worst state described by the measure. The anchoring process uses a function of $\beta_{ij} = \beta'_{ij} \times w_{cardinal}/w_{ordinal}$ to rescale each DCE latent coefficient, where $w_{cardinal}$ is the worst latent cardinal value and $w_{ordinal}$ is the equivalent ordinal health state value. The second option, similar with hybrid model, predicts the statistical relationship between ordinal latent and cardinal health state values using observed values for 'pits' states. The function could be specified as Formula 5

$$cardinal\ value_j = f\left(ordinal\ latent\ value_j\right) + \varepsilon_j \qquad \text{Formula (5)}$$

where $cardinal\ value_j$ was the mean utility value for ‘pits’ health state j derived by cardinal method, and $ordinal\ latent\ value_j$ was the latent ordinal utility. Then the mapping function can derive the Ordinal Least Square (OLS) or Maximum Likelihood (MLE) relationship and generate other state values with ordinal latent values. A mapping method considers more information than the ‘anchoring with worst’ method, if the necessary cardinal values available.

4) Anchoring with duration (also named DCE_{TTO}): This format involves presenting participants with two different health states, each defined by a specific preference-based measure, and attaching a duration to these states as the expected time spent on this state before death[122]. The data analysis is modelling interacted ordinal data to derive marginal effects. α_i is the left-right or other constant effects. β'_2 is the marginal contribution of level factors X_{ij} at time level t_{ij}. With the assumption of linear time preference, the partial derivative with respect to time, β_1, is used as the anchored coefficient[44].

$$\mathrm{U}_{ij} = \alpha_i + \beta_1 t_{ij} + \beta'_2 X_{ij} * t_{ij} + \varepsilon_{ij} \qquad \text{Formula (6)}$$

Notably, over 70% of anchored DCE health state valuation studies adopted this format[123]. With varied numbers of attribute and survey questions, empirical

evidence supports the reliability, consistency, individual-level congruence and cardinal value robustness of DCE_{TTO} data[124] (an example presented Figure 1a, Appendix A).

5) Anchoring by two-way triplet design with perfect health or death: this design aims to provide a triplet ranking information with two separate pairwise choice tasks, with which the respondents have coherent decision-making process[86]. The first choice task asks respondents to compare two different impaired health states with same 10-year duration (health state A and B comparison). If the third comparison is with perfect health, the second task compares one impaired health state with longer duration level against a perfect health state with shorter duration level [125] (health state B and C comparison), building upon the TTO decision-making framework. If the third comparison is with death, the second task compares impaired health state with 'immediate death' as Health state C[126]. Examples for the two-ways method are listed in Figure 1b and 1c, Appendix A. The original research did not present anchoring function but noted the methodological feasibility for anchoring.

   This format elicits direct preferences through comparing with perfect health or better-or-worse-than-death inquiries, bypassing the need for ordinal data ranking[127]. However, there's a noted concern regarding the potential breach of random utility theory when ordinal data is used to rank health states with or without death[120]. Another notable drawback is that this method requires participants to make twice as many decisions as the DCE_{TTO} format, potentially increasing respondent burden.

6) Anchoring with both duration and death information: this method presents one combined question involving three health states (imperfect health states A and B, and immediate death C). Respondents are asked to identify the best and worst states following the attribute information. This format diverges from the third by allowing variable durations for states A and B, compelling a consideration of health-time trade-offs. A significant benefit of this design is the detailed death ranking and duration data it yields. This format relaxes the transitivity requirement seen in the two-way triplet design for creating ordinal data[128], where a complete ranking of all four options provided, assuming that respondents apply a consistent decision-making approach to both best and worst choices (Figure 1d, Appendix A). Norman *et al*., introduced a function to explicitly consider death with duration. The added term $\beta_i D$ is a dummy-coded variable capturing the latent utility of death[129].

$$\mathrm{U}_{ij} = \beta_i D + \beta_1 t_{ij} + \beta_2' X_{ij} * t_{ij} + \varepsilon_{ij}$$ Formula (7)

1.4.7 Why consider DCE for health state valuation?

Health economists and HTA bodies have favored cardinal scaling method SG and TTO for a long time[4, 41], but the use of DCE is gaining preference over traditional methods. This shift can be attributed to several advantages that DCE offers.

The first advantage is reduction of bias. Cardinal methods are susceptible to various biases including utility curvature, probability weighting, loss aversion, and scale compatibility[130]. For instance, TTO assumes linear utility over time, which may not reflect real-life preferences that can fluctuate at different life stages. DCE, by contrast, can mitigate these biases[131]. DCE valuation design can be designed to prevent utility curvature, probability weighting and loss aversion (with no duration attribute)[130].

The second advantage is lower cognitive burden. Cardinal methods often require respondents to give a score health states, which can be challenging, especially for individuals with lower educational attainment[4, 67, 132]. The ordinal method is based on health state rankings achieved by multiple choices. Salomon et al. noted that DCE had a lower cognitive burden and less abstract than TTO for diverse populations[133]. Bansbacka *et al.,* noted in their study that TTO was too cognitively demanding to be consistently completed by young respondents[44]. A health-state valuation method comparison study with five attribute vignettes using VAS, TTO, and DCE find that DCE has the highest feasibility and test-retest reliability[134]. DCE would has the potential to present more information at the same time.

The third advantage is its ease of administration. DCE aligns well with the structure of PBM, as each health state in a PBM is composed of different attribute levels. The PBM classification system accords with the requirement of ordinal scenario design. This compatibility makes DCE easier to administer in comparison to methods like TTO[135], which require different task forms for health states better or worse than death[61]. DCEs can be efficiently conducted via online surveys, or mail, reducing resource requirements and complexity.

DCE is increasingly becoming the preferred method for eliciting ordinal preferences in health economics[107], particularly in studies with a large number of attributes (See Chapter 3 for more information)[123]. Researchers have demonstrated the feasibility of estimating value sets for lengthy PBMs using DCE, as seen in studies employing PROMIS-29 and ASCOT+EQ-5D measures[134, 136]. Despite these advantages, it's important to note that there is no consensus on the best practice for DCE study design.

Often, ordinal questions are used as warm-up exercises rather than as a means to generate population tariffs, and more evidence is required to establish the most effective DCE methodologies[67].

1.5 Conclusion

The market failure in the healthcare market requires health technology assessment institutes to use economic evaluation to inform the unavoidable decisions in health care. PBMs have been developed to generate QALYs. Researchers valued the quality adjustment weight using cardinal (TTO, VAS, SG) or ordinal methods (ranking, BWS, DCE). Cardinal methods have long been regarded as the preferred methods for generating value sets for PBMs and recommended by valuation protocols. However, cardinal methods can never be regarded as perfect, and the use of ordinal data is especially attractive for long measures (attribute number >13, where most of the PBMs are described by $\leqslant$13 attributes. Among the ordinal methods, DCE is a promising technique of health state valuation with lower cognitive burden, reduced bias and more flexible application, especially with long PBMs. Chapters 2 and 3 will introduce the novel EuroQol Health and Wellbeing instrument, why an ordinal method for valuation may be desirable rather than a cardinal method, and what are the specific DCE design strategies available.

Chapter 2 From EQ-5D to E-QALY: Beyond Health

Chapter 1 explored the need for economic evaluation in the health sector, delving into both cardinal and ordinal methods for eliciting the value of health states. Recognizing that quality of life encompasses various domains, it is typically characterized using PBMs. This chapter presents the most commonly used PBMs globally, EQ-5D, which concentrates on assessing changes in physical and mental health, and its relationship with the novel EuroQol Health and Wellbeing (EQ-HWB) measure, designed to evaluate more extensive health and wellbeing changes.

2.1 The health QALY and its development: take EQ-5D as an example

2.1.1 Descriptive system and questionnaire

The EQ-5D is a standardized Generic Preference-Based Measure (GPBM), measuring, comparing, and valuing health states across diseases and informing resource allocation. EQ-5D covers five dimensions: *Mobility, Self-care, Usual Activities, Pain/Discomfort, and Anxiety/Depression* (Table 2). Among the dimensions, Mobility, Self-care, Usual Activities, Pain/Discomfort are physical health dimensions, with one dimension Anxiety/Depression evaluates mental health[52]. The three-level, EQ-5D-3L, has three response levels of severity (no problems, some problems, and extreme problems/unable), defines 243 health states. However, substantial empirical evidence suggested the three-level instrument had ceiling effect and was insensitive with slight health decrement or improvement[137-139]. EQ-5D developer EuroQol developed the new five-level (with five levels: no problems, slight problems, moderate problems, severe problems and extreme problems) version, EQ-5D-5L. The new measure generates 3125 health states. Initial evidence suggested that the new measure reduces the ceiling effect problem to some extent[4].

The EQ-5D consists of two parts: the first part is a descriptive system, designed to evaluate the respondent's current situation (ticking the one box that best describes health TODAY). The second part is a visual analogue scale asking respondents to report their health on a 0 to 100 scale. Respondents can either self-report with support from trained researchers, or proxy reported by an informal caregiver or relative in limited situations. Then the health state utility score can be calculated using scoring algorithm.

Table 2 EuroQol generic preference-based measures

	Target Population	Quality of Life	Dimensions	Attribute number	Levels	States
EQ-5D	Patients	HRQoL	Mobility, Self-care, Usual activities, Pain/discomfort, Anxiety/depression	5	3 or 5	243 or 3125
EQ-HWB	Social care users, Patients, Carers	Social-care related QoL, HRQoL, Carer related QoL	**Activity:** Vision, Hearing, Day-to-day activities, Self-care, Mobility; **Physical sensations:** Pain, Pain severity, Discomfort, Discomfort severity, Sleep problems, Energy; **Relationships:** Loneliness, Support, Relationships; **Cognition:** Memory, Cognition Feelings and emotions: Sad, Hope, Anxiety, Safety, Anger **Self-identity:** Self-respect; **Autonomy:** Control, Coping, Autonomy;	25	5	5^{25}
EQ-HWB-S	Social care users, Patients, Carers	Social-care related QoL, HRQoL, Carer related QoL	Mobility, Daily activity, Exhaustion, Loneliness, Cognition, Anxiety, Sadness/ depression, Control, Physical pain	9	5	5^{9}

The descriptive system, named EQ-5D, was sent to EuroQol members in 1996 to define key information in the questionnaire and prepare for global use[140]. A plethora of research assessed the acceptability, feasibility, reliability, validity, and responsiveness of EQ-5D, through its 45 years of history. Studies provided the acceptability and validity evidence in different populations and settings, including depression, diabetes, rheumatoid arthritis, skin conditions, cancer and cardiovascular disease [52, 141]. However, there are also mixed evidence indicating the validity for use in some diseases or conditions[4]. The EQ-5D-5L value sets has been published in the UK recently, but the UK HTA institute NICE evaluate the value set with cautious[142]. To satisfy the demand of evaluating child health, a age-specific version EQ-5D-Y instrument was developed[143]. As an international health measure, the EQ-5D has been translated into 169 languages[144] and value sets are available in 25 countries[61] and more studies are still waiting for publication.

2.1.2 Valuation and application

An important point to remember is that the EQ-5D valuation methodology has been iteratively tested and modified in the past 20 years. Different methods yield different value sets for the same instrument. A starting point was the Measurement and Valuation of Health (MVH) study conducted by the University of York in the 1990s{Williams, 1995 #785;Kind, 1998 #943}. The early version of the EQ-5D UK value set was based on TTO survey data from UK general population, and each respondent needs to value 45 health states[146], followed by VAS valuation tasks. EuroQol encouraged researchers to adopt same valuation methodology for generating the country-level tariff, to increase the comparability and reliability. However, the protocol applied in different forms, endorsing various valued health state selection criteria and the data exclusion rule for EQ-5D-3L[147], where a minority of countries tested the feasibility of using ordinal data to generate the EQ-5D-3L value set.

After developing the EQ-5D-5L measure, the EuroQol group developed an official valuation protocol, suggesting using conventional TTO for health states better than dead and lead-time TTO for states worse than dead[1, 53]. The computer-assistant version has been called EuroQol Valuation Technology (EQ-VT) protocol, which consists of a web-based data collection system, Quality Control (QC) tool, tracking tool to know the progress of study and a parsing tool for data analysis [148]. EQ-VT data collection enables TTO data to be combined with DCE data to generate the value set[119], though there is still the option of producing a TTO-only value set [177]. The preference elicitation derives from a strong assumption that the utility difference

resulted from health state difference, instead of left-right preference and lexicographic preference with health[149-151]. Each respondent needs to answer 10 TTO tasks and 7 DCE questions, and the valuation has quality control throughout the data collection process. Age, education, gender and ethnicity are the main characteristics of quota control [152].

The EQ-5D-3L/5L utility functions were specified as additive models with 10 coefficients (3L) or 20 coefficients (5L), constructed with the homogeneity TTO+DCE hybrid method and linear additive utility function. As introduced above, the mixed logit model relaxed the preference homogeneity assumption estimating a distributed preference weight, or use heteroscedasticity model with an assumption that the more severe health states tend to have larger variance by its utility function[52, 153, 154]. A EQ-5D-5L value set may also generated through crosswalk: a mapping function can be generated by collecting the EQ-5D-5L and EQ-5D-3L responses from a smaller sample size than valuation study[155].

2.1.3 Criticism for EQ-5D measures

EQ-5D is the most popular GPBM, but it needs to note that EQ-5D is not a perfect measure due to the limited dimension and level number. Although some evidence indicated that GPBMs were able to evaluate the health impact of depression and anxiety[156], a study with Finnish population showed that EQ-5D index were not sensitive for delusional or bipolar I disorders[157]. UK population survey with hearing imperfect population found a similar outcome[31], which was also predictable for vision. Other populations where the EQ-5D did not show a statistically significant reduction were long-term QoL improvement with care, multiple sclerosis symptoms and schizophrenia[158]. Such evidence posed challenges for reflecting the utility change of patients.

EQ-5D was criticized for ignoring non-health benefits, such as wellbeing and capability [159]. EQ-5D was more suitable for informing healthcare resource allocation decisions on pharmaceutical drugs and medical instruments than informing decisions with social care (e.g., care for independent senior people) and palliative care[36]. The content of EQ-5D was developed through literature review and expert judgement 20 years ago. One wellbeing attribute social functions and wellbeing was combined into usual activities in the later stage of EQ-5D development[52].

Practically, short measure focusing on health change poses less conceptual and valuation challenges. However, there is interest among policy makers and clinical professionals in measuring changes in wellbeing[159]. To achieve the cross-disease

comparison between traditional healthcare interventions and care and evaluate broader health improvements, health economists are interested in developing a new measure that reflects changes beyond the dimensions of physical and mental health covered by EQ-5D[144].

2.2 The health and wellbeing QALY and its development: EQ-HWB

This section provides a more detailed introduction for EQ-HWB measure and related researches: how does the measure developed through E-QALY project, health and wellbeing concepts in EQ-HWB and the progress of EQ-HWB related research[160].

2.2.1 Extending the QALY project and EQ-HWB

EQ-HWB was developed by the Extending QALY (E-QALY) project, which`- was led by the Sheffield Centre for Health and Related Research (ScHARR) and in collaboration with other institutions and the EuroQol Group[160]. The aim of E-QALY was to 'develop a broader generic measure of health and wellbeing for use in economic evaluation across health, social care, informal care and public health based on the views of users and beneficiaries of these services, including informal carers'[161].

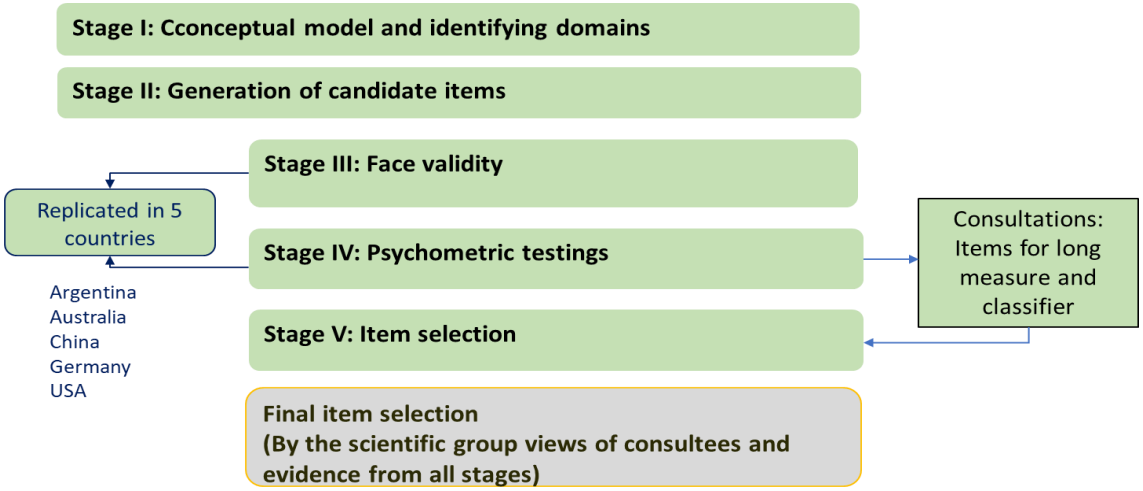
The project has six stages (Figure 2): identify potential domains, generate items and questions for each domain, test the face validity of items with representatives, psychometric testing of selected items with users, item selection, and the valuation, implementation and assessment[161].

The first stage was to identify potential domains by literature review. The starting point of this work was a targeted review of the qualitative evidence on the 'impact on QoL' by selected population groups of patients, social care users, and informal carers[162]. The analysis used a modified Wilson and Clearly framework, which presented an analysis framework for correlation between circumstance / biological / physiological variables and quality of life, to organize and analyse the extracted factors. The identified themes and sub-themes were then selected based on their importance, feasibility for self-reporting, non-instrumental consequences and non-overlapping concepts[39]. After the dropping and merging of sub-themes, the review process resulted in 7 themes (domains) and 26 sub-themes (sub-domains) covering both health and wellbeing identified in this stage (Seven themes are: Feelings and emotions, Cognition, Self-Identity, Autonomy, Relationships, Physical sensations, Activity)[163].

The second stage was to generate potential item questions for domains and sub-domains. An item pool was developed from a review of 30 existing measures, data banks and de novo from the qualitative review. Item inclusion was based on item

selection criteria as follows: ease of completion, avoid items that are value-laden, coverage of sub-domains and severity range, coverage of current QoL measures, translation and localization suitability, and suitability for valuation[164]. 687 potential items were identified from the review and 89 items were left after selection. There were also considerations on the items related to more than one domain, and the recall period as well as negative or positive wording to make the whole item pool consistent[165]. For those sub-domains without potential items, new items were generated. After a review of items from stakeholders, advisories, Patient and Public Involvement and Engagement (PPIE) group and necessary refinement of items, 8 items were generated or added to the 89, and 97 items remained. The measures have experimental status which allows items to be refined as they are tested.

Figure 2 overview of E-QALY project



Source: John brazier. Sheffield mini master class PowerPoint: https://www.youtube.com/watch?v=KTlsIvqyhNI&t=2337s and Brazier J *et al.,* (2022).

Stage three was a qualitative multi-national semi-structured interview to test the face validity of the 97 items from stage two. 168 participants, including social care users, patients, mental health service users, carers, and healthy people, were one-to-one interviewed in the UK, Argentina, Australia, China, Germany, and the USA. Each respondent was presented with around 40 items from two to three domains. They were asked whether they had a preference and alternative wording if they did not like the item. The items were selected based on the meaning /interpretation of meaning, positive or negative comments (though not transcribing the verbal recording into qualitative analysis documents), and suitability of response. The result shrunk the item pool to 64 items (3 new items added)[165].

The fourth stage was to assess the psychometric performance of the proposed items through a large survey. 4830 participants, covering patients (acute, long-term

conditions, and mental health service users), social care users, carers and the general public, completed the large-scale psychometric survey among the six countries. The respondents completed E-QALY items, EQ-5D, Short Warwick Edinburgh Mental Wellbeing Scale (SWEMWBS), and social care measure (ASCOT) to test the distribution of responses, domain structure, item performance (using Item response theory) and construct validity of the proposed 64 items[166]. Dimensionality and its validity in multicultures were explored by exploratory and confirmatory factor analysis. Item performance was tested to check its ability on discriminating groups, as well as the rate of missing data and floor or ceiling effects. Views of the advisory group were included in the selection of items in stage 3. 32 items performed well, 25 had mixed evidence and 7 performed poorly[161].

Stage five involved the selection of the items for EQ-HWB and short classifier EQ-HWB-S. The E-QALY group aimed to have at least one item for each sub-domain. This item selection was completed by collecting evidence from stakeholders including the project advisory group, PPIE group, HTA institutes and EuroQol Group membership. For the long version EQ-HWB, stakeholders (academics, those working in HTA and in pharmaceutical companies) were invited to provide opinions with presented items: to include, reject or undecided about the item in an online survey. Top-ranking item in each sub-domain were retained. For the EQ-HWB-S, a second round of consultation with 71 experts was conducted, with two selection concerns that the EQ-HWB-S items should be core items and correlation among the sub-domains should be acceptable. Respondents were asked to rank the items and indicate whether they strongly recommended, recommended, or were not sure or do not recommend the item as suitable to inform decision-making in the context of economic evaluation. 7 of 10 most highly ranked items were selected and two items were added for the draft EQ-HWB-S. The research team also considered a pilot interview evidence with DCE and TTO method, to test the feasibility of valuation of each item[161]. It resulted in a change of the coping item to a control item. The final version of EQ-HWB had 25 items and EQ-HWB-S had 9 items with different response options and a seven day recall period (see in Table 2 and 3, Appendix A) [161].

Stage six is the valuation of the new measure. For the shorter classifier EQ-HWB-S, a feasibility study with 520 UK participants tested valuation with EQ- Portable Valuation Technology protocol (version 2)[119]. Each respondent participated in an online videoconference interview to complete a questionnaire with 4 practice TTO questions + 7 TTO questions + 7 paired comparison DCE questions[167]. The result indicated that participants understood the TTO and DCE tasks as was expected[167]. Qualitative

evidence from the population group with high education level also proved that TTO questions were fairly or very easy to understand but held dissent opinion on what other information should be added to enrich the measure[168].

Apart from efforts to generate a value set for the new measures, there is emerging evidence on the validity and efficiency of the new instruments in different populations, including the general population[169], care givers[170], and parents[171].

2.2.2 EQ-HWB

The EQ-HWB definition of health and wellbeing is characterized as a bottom–up process, in favor of a more flexible, inclusive dimension structure that focused on health, social-care and carer related quality of life. Wilson and Cleary's model for health-related quality of life, has been adopted to identify the main dimensions [64], incorporating bidirectional relationships and interactions between health symptoms and emotional feelings [46]. Other sources of HWB are literature review of health and wellbeing PBM measures [65], and qualitative interview data with future measure users and stakeholders, especially the UK HTA institutes and social carers[39, 166, 172, 173].

The EQ-HWB covers seven domains (physical sensation, feelings and emotions, activity, self-worth, control and coping, relationship, cognition) [161]. The 25 items of EQ-HWB are: *Vision, Hearing, Mobility, Daily activities, Self-Care, Sleep, Fatigue, Loneliness, Support, Memory, Concentrating/ thinking clearly, Anxious, Unsafe, Frustrated, Sad/depressed, Hopeless, Control, Coping, Stigma/belonging, Enjoyable activities, Self-worth, Pain severity and frequency, Discomfort severity and frequency*. Each sub-domain identified in the E-QALY stage is represented by one item, with the exception of pain and discomfort. Loneliness, stigma, and concentration have one item each and there is no item specifically for dignity. Stigma/belonging, Enjoyable activities, Self-worth were positively worded while the rest were negatively worded. Each item is designed with five response levels.

The EQ-HWB-S, with nine core attributes, represents a more concise variant of the full EQ-HWB measure that was amenable to valuation. Items included are *Mobility, Daily activity, control, concentration, anxiety, depression, loneliness, fatigue, and pain severity.*

2.2.3 Valuation

The EQ-HWB-S is amenable to valuation using established techniques. A pilot valuation study in the UK, following the EQ-VT protocol, was conducted to test the validity of the existing valuation method for EQ-HWB-S[168]. An alternative method,

called Online elicitation of Personal Utility Functions (OPUF), has recently been used to create the value set of UK and Australian general public population, and German rheumatic disease and diabetes patients. All of the published valuation study targets on testing a promised EQ-5D valuation method using the new measure (until December 2023). The two studies proved that the health state valuation theory and approaches could be applied to measures extended beyond the scope of physical and mental health. The additional dimensions of EQ-HWB-S, compared with EQ-5D (9 vs. 5) provided respondents with more detailed scenarios, though the increased amount of information could potentially raise the difficulty of the task. Respondents understand the classic valuation designs (DCE and TTO) well and the data quality is comparable to the EQ-5D health state valuation data (OPUF){Schneider, 2022 #849}. However, other valuation approaches (e.g., generating a value set solely through ordinal data) has not been fully tested with the EQ-HWB (-S) measure. To present an overview of other potential valuation methods, this study conducted a literature review and other promising alternative methodology options will be introduced in Chapter 3.

As EQ-HWB-S served as a new measure reflecting broader health change, some of the ongoing valuation studies explored the utility influence of using a “bolt-on” version of EQ-HWB-S. No result has been published yet (until November 2024).

Currently, there is no specific valuation protocol tailored for the full EQ-HWB measure. While the EQ-VT protocol originally designed for valuing the EQ-5D health measure, concerns were raised about its application with the full EQ-HWB measure. The extensive information presented in the valuation tasks for the EQ-HWB measure would require that careful consideration must be given to the selection of the valuation method.

For this valuation research, the EQ-HWB items were classified into physical health, mental health and wellbeing to distinguish across them:

Physical health: *Vision, Hearing, Mobility, Daily activities, Self-Care, Discomfort severity, Discomfort frequency, Pain severity, Pain frequency.*

Physical but influenced by mental health: *Sleep, Fatigue, Memory, Concentrating/ thinking clearly.*

Mental health: *Anxious, Unsafe, Frustrated, Sad/depressed, Hopeless.*

Wellbeing: *Loneliness, Support, Control, Coping, Stigma/belonging, Enjoyable activities, Self-worth*.

2.3 Conclusion

The EQ-HWB and EQ-HWB-S represent a significant advancement beyond the EQ-5D, covering aspects of health and wellbeing identified as important by service users and informal carers. The EQ-HWB is not simply an update or extension of the EQ-5D, but rather a novel measure designed to capture a broader range of benefits, covering both health and wellbeing. It capture a broader notion of change of health-, social care- and carer-related quality of life. The EQ-HWB therefore offers a more inclusive evaluation of outcomes in CEA studies and multi-disciplinary comparison in policymaking.

In the perspective of valuation, the 9-item EQ-HWB-S measure is more feasible in terms of length and can be valued using the EQ-VT protocol. However, the EQ-HWB-S does not contain the full richness of wellbeing themes and sub-themes captured in the EQ-HWB which potentially limits what it is able to capture. This raises the issue of whether it is possible to value a wider set of items from EQ-HWB to enable the wider set of aspects covered in EQ-HWB to be reflected in utility values. Balancing comprehensiveness in the state description with practicality remains a key challenge in developing effective tools for evaluating health and wellbeing outcomes within a QALY framework. In the next section, I will introduce some feasible valuation designs specifically for long measures.

Chapter 3 Literature Review of DCE Health State Valuation Studies

The preceding sections have provided an overview of health state valuation methodologies and a brief comparison of EQ-5D and EQ-HWB. Recently, health economists have increasingly adopted the DCE approach[174], and as noted by Carson and Louviere[80], DCE is a 'general preference elicitation' survey approach asking respondents to choose between two or more alternatives, where at least one attribute is 'systematically varied'.

The initial application of DCE in measuring health state preferences was reported by Hakim and Pathak in 1999[82]. Since then, DCE has been widely used in health state valuation research and is recommended in the EQ-5D-5L valuation protocol[53, 84, 175]. This section included literature from September 2018 to December 2022 in both English and Chinese language. It highlights recent advancements and trends in DCE methodologies within health state valuation and summarizes viable study designs for long measures. The paper has been published, available at:

https://doi.org/10.1007/s40258-023-00794-9. This is an extended version of the published review.

3.1 Published evidence

Mulhern *et al*. and Bahrampour *et al*. conducted two comprehensive review studies focused on DCE valuation of PBMs [84, 175]. The first study, conducted by Mulhern et al., spanning from 1999 to May 2018, identified 63 studies satisfying the inclusion criteria, generally exhibiting high quality. In contrast, the second study by Bahrampour *et al*., covering the period until 2018, differed in its search terms by not including BWS Case 3 and excluded simulation data studies. This resulted in the identification of 38 studies. Like the first study, the quality of studies were high, with an average score of 86.5%. Both studies provide valuable insights into the application and quality of research in the field of health state valuation using DCE and PBM methodologies.

The review of DCE studies in health state valuation revealed notable trends and methodologies. There was a noticeable increase in the number of DCE studies, with 57% of the reviewed papers published in the last three years of the review period (2016-2018). Over half of the studies created a value set, while the rest of them focused on study design strategies, preference heterogeneity, and data modelling comparisons.

EQ-5D measures (3L, 5L, and EQ-5D-Y) were the most commonly valued PBMs across these studies. Most studies employed digital DCE surveys. The sample sizes

varied, with half of the studies involving over 1000 participants, while two studies had fewer than 100 participants. The choice tasks in these studies typically featured five to six attributes, aligning with the attributes of popular PBMs like EQ-5D or SF-6D, with a range of 4 to 20. The majority of studies presented respondents with over ten choice questions, with a range from 4 to 32. D-efficiency design was prevalent, and priors often derived from value sets of other countries or from a small-sample prior survey. The conditional logit regression model was the most common analytical approach[176] for anchoring.

To manage the complexity of tasks with numerous attributes, researchers employed various presentation formats, such as coloured balloons for levels or highlighting with different colours, to reduce cognitive burden for respondents. It was found to improve response rates to some extent.

In conclusion, valuation studies employing the DCE method have garnered significant attention. However, it is important to note that an universally accepted "gold standard" for study design and the valuation of long measures remains unclear. This review updated the evidence, including study design, information presentation, anchoring strategies and the data analysis methods, and tried to summarize all of the feasible study design strategies for generating a value set with long measures.

3.2 Review question and methodology

3.2.1 Literature Search

This literature review was the first review of this literature covering both English and Chinese database. Recently, DCEs gained prominence in generating health state utilities within China. A notable instance includes its application in deriving the Chinese value set for the SF-6Dv2 measure[177], but there has been an absence of systematic searches and review of such studies. To formulate our search strategy, this study referred to prior systematic reviews [84, 175] and translated the English keywords into Chinese. The searched English databases were PubMed, Cochrane, and Chinese databases Wanfang and CNKI.

Our search strategies were developed based on existing published evidence. While Bahrampour *et al*. represented the most recent review in terms of publication date[175], the registration details on the PROSPERO website suggested their literature search concluded earlier. Additionally, Bahrampour *et al*. excluded studies that utilized second-hand data, contrasting Mulhern's inclusion of studies with both primary and secondary data[84, 175]. Our review expanded upon the foundation laid by Mulhern et

al[84], aiming to include a broader range of evidence. The search terms encompassed descriptive keywords related to discrete choice survey (e.g., "discrete choice experiment," "DCE"), health state valuation (e.g., "value set generation"), and Multi-Attribute Utility Instruments (e.g., "preference-based measure," "PBMs," "MAUIs," "EQ-5D"). A scoping review was conducted to identify the various terminologies and Chinese translations for DCE methods (such as "paired comparison," "case 3 Best-Worst Scaling"), and an external Chinese expert was consulted for reviewing the search terms. In addition to the term ordinal method, some institutions or researchers use conjoint analysis to represent the choice methods or specifically for DCE. The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) defined conjoint analysis as a concept involving ranking, rating, or multi-attribute choice design for general health preference evaluation[178]. However, Louviere *et al* (2010) reported that the conjoint measure theory was about the factorial manipulations of levels that applicable for utility measurement[179]. This research included “conjoint analysis” as the English search term.

The comprehensive list of English and Chinese search terms is detailed in Table 4, Appendix A. The initial English search was completed in March 2021 and updated in December 2022 to include the most recent papers. The Chinese database search was completed in early October 2022.

3.2.2 Inclusion and Exclusion criteria

Health state valuation or methodology study papers were included if the study used DCE design or paired Case 3 (multi-profile case) BWS to generate a value set for a PBM. Case 3 BWS asked respondents to select the best and worst scenarios with more than one multiple-attribute profiles where the choice experiment was in line with DCE [180, 181]. Papers were excluded if:

1. Only non-DCE methods were used[72]*;*

2. DCE studies targeting a monetary parameter ratio or Willingness-To-Pay (WTP) for a certain intervention.

3. Studies valuing partial health states where not all attributes were considered, or health states that were not derived from Preference-Based Measure (PBM), where a value set cannot be developed.

4. Quantitative studies using DCE but not reporting the statistical analysis results, reviews and qualitative studies.

5. Papers published before September 2018.

6. Data generated from software simulation instead of real-world survey, or where the study design is not reported, conference abstracts where full text was not available, and replicated articles in various languages were excluded.

HW and external expert LJ designed and translated the search terms. HW reviewed article titles and abstracts. The full text of remaining articles was reviewed by HW. PhD supervisors DR and JB finalized the included articles. DR and JB provided suggestions for developing the data extraction sheet to confirm that the extracted information was consistent with the review objective. HW extracted the information and DR assessed the data extraction quality for a subset of articles.

3.2.3 Data Extraction

Data extraction was conducted using a designed data extraction sheet. The extraction sheet comprised four key components: 1) General study information, including sample details, measurement, and data characteristics; 2) Study design elements, such as attributes and levels, attribute categories, the number of scenarios and choice sets, anchoring methods, the questions posed, and the statistical analysis approach; 3) analysis specifics, focusing on whether the results were latent or anchored, and the logical consistency of the findings; and 4) reported research limitations and recommendations, particularly concerning the methodological choices in DCE. The information in data extraction sheet was based on insights from previous reviews [84, 175]. Key information was reported for the main study characteristics, study design and analysis. Trends since the previous reviews were outlined in the discussion.

3.2.4 measure concepts

In this section of our study, three distinct measure concepts to categorize and assess the effectiveness of the identified evidence are used: wellbeing measures, health measure and long measure. The criteria for defining wellbeing measures are grounded in three consensus points: 1. Recognition of the five wellbeing measures (GHQ-12, WEMWBS, ONS-4, ICECAP-A, ASCOT) commonly utilized in the UK, as referenced in a previous study[172]; 2. Inclusion of measure attributes that assess life pleasure, attitude, and qualities, aligning with the principles of hedonism theory; 3. Identification of a measure as a "wellbeing measure" either by the original study authors or through the predominance of "wellbeing" attributes within it. A health measure is defined as “focuses on health attributes but may also encompass aspects of wellbeing”.

3.3 RESULTS

3.3.1 Identified studies

The search identified 1133 English language records and 46 Chinese language studies using DCE and preference elicitation search terms, where 1172 articles were included after duplicate checks. 1106 records were from PubMed, 20 articles from the Cochrane database, 16 articles from Wanfang, and 30 from CNKI. All of the studies reported DCE study design and no case 3 BWS articles were identified. After screening titles and abstracts, 1063 articles were excluded, leaving 109 articles. The assessment of full articles further excluded 44 articles. There was one Chinese language article, and 64 English articles satisfied the inclusion criteria (Figure 2, Appendix A). All the 65 papers (Table 5, Appendix A) were double-checked to ensure no research study overlapped.

3.3.2 General characteristics

An increasing number of works were identified in the reviewed years in comparison to the reviews of the literature up to 2018[84, 175]. More papers were published in 2021 (n=19) than 2020 (n=17) and 2019 (n=9) but cannot easily be compared to 2022 (n=15), since this is not a full calendar year. Overall, the majority of the studies were conducted in Organization for Economic Co-operation and Development (OECD) countries, examples including the UK (n=13), the USA (n=6), Australia (n=6) and Netherland (n=7). Other countries with more than one research identified were Germany, China (n=4 for each country), Italy (n=3), Canada, France, Poland, Spain, Hungary and Slovenia (n=2 for each country). Denmark, Egypt, Ethiopia, Malaysia, Mexico, New Zealand, Japan, Peru, Portugal, Russia, Slovenia, Tunisia, Philippines and Thailand all provided one. Compared with published reviews in 1999-2018, there was an increase in studies coming from ‘developing’ countries (15 % in 1999–2018 compared with 25 % in 2019–2022), but the UK, the USA, and the Netherlands were still top four publication sources.

3.3.3 Sample size and PBM measures

Most studies (n=60) sampled the general population, and stratified the respondents by gender, age, educational level and region. Other studies collected data from adolescents[182], parents[183], diabetic macular edema patients[184], elderly people [125, 185] or people with hemophilia [186]. The sample size varied among the studies. Forty-nine studies (46 valued by general public and 3 valued by a specific group) interviewed over 1000 respondents, with a sample size ranging from 220 to 13623 (Table 3). The average sample size was 1704, with a number for the general population of 1948 samples and for the specific population of 797.

The proportion of studies administrated online was similar with the previous review. 37 studies (60%) collected data online with an online panel. In comparison, Mulhern et al [84] identified 37 (59%) of all the papers employed online administration mode. Out of the 25 off-line studies, 21 employed software-assistant data collection. Two studies used mixed data collection strategy and one study did not mention their data gathering method (Table 3).

Table 3 Experiment design characteristics

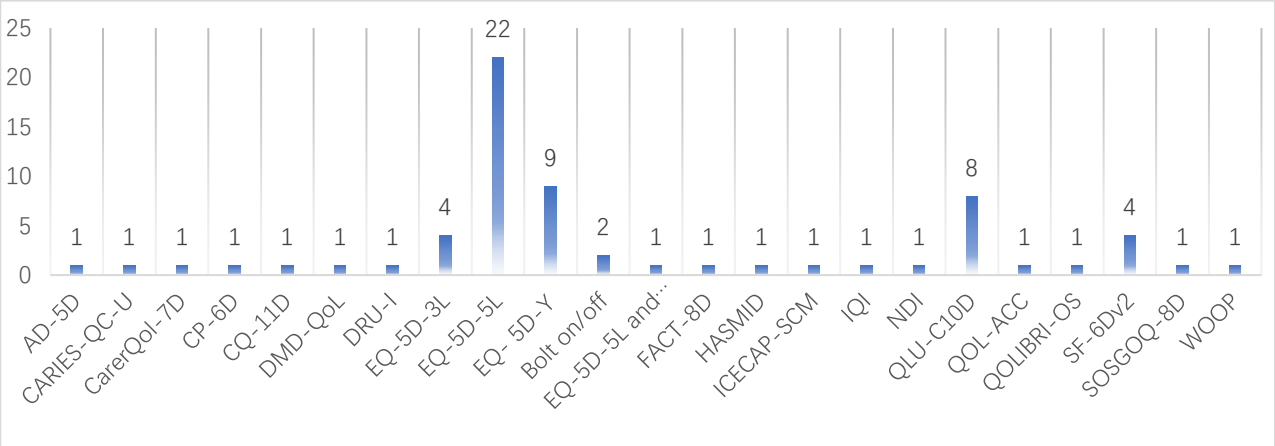
Characteristics	Level	Identified studies
Attributes Number (Range: 5~13)	**Range: 5~13**	
	5	38
	6~7	10
	8~10	6
	11~12	10
	13	1
Number of Levels	3	15
	4	11
	5	34
	6	5
Anchoring	Anchoring with cTTO data	24
	Anchoring with duration	29
	Anchoring with VAS data	3
	Others[1] (rescaling)	5
	No anchoring	4
Choice set	Pairs	62
	Triplets[2]	3
Choice tasks per participant	**Range: 7~28**	
	≤12	38
	13~28	25
	Not mention	2
Survey mode	Interview offline	25
	Interview online	37
	Mixed	2
	Not mention	1
Question asked	Prefer	32
	Better	10
	“Which one do you pick”	4
	Following EQ-VT (v1/v2 question format mentioned)	14
	Not mention	5
Total number of Choice sets	**Range: 28~960**	
	≤120	20
	121~196	26
	197~960	16
	Not mentioned	3
	In total	65

Note: 1. Re-scaling anchoring option includes re-scale with existing tariff or re-scale with the minimum/maximum utility value.

2. All of the studies included used death as the third state.

EuroQol HRQoL measures (5D-3L, 5D-5L and 5D-Y) were the most commonly valued PBMs. The EuroQol international protocols for valuing EQ-5D-5L[53] recommend using TTO and DCE meant there were 21 papers generated EQ-5D-5L value set under the recommended framework. Nine studies generated EQ-5D-Y value sets. Other studies measured generic PBMs including EQ-5D-3L (n=4), EQ-5D bolt-on/bolt-off measures (n=2), SF-6Dv2 (n=4), EQ-5D-5L plus ASCOT (n=1), the informal caregivers' life quality measure CarerQol-7D (n=8) and infant health-related quality of life instrument measure IQI (n=1). Various Condition-Specific PBMs were also valued, such as the European Organization for Research and Treatment of Cancer (EORTC) cancer utility measure instrument EORTC-QLU-C10D[187]. (n=6), the impact of self-management on quality of life in diabetes measure HASMID, a tool for palliative and supportive care ICECAP-SCM [188], diabetic retinopathy measure DRU-I[189], traumatic brain injury outcome measure QOLIBRI-OS[190], Alzheimer measure AD-5D[135] and cerebral palsy measure CP-6D[135, 191] (Figure 3).

Figure 3 Measures used in identified articles



AD-5D: Alzheimer's Disease Five Dimensions; CARIES-QC-U: Caries Impacts and Experiences Questionnaire for Children; CarerQol-7D: Care related quality of life-7 dimensions; CP-6D: cerebral palsy quality of life-6 dimensions; CQ-11D: Chinese medicine quality of life-11 dimensions; DMD-QoL: Duchenne muscular dystrophy quality of life; DRU-I: Diabetic Retinopathy Utility Index; EQ-5D-3L/5L: European Quality of Life 5 Dimensions 3 Level/5 Level Version; EQ-5D-Y: European Quality of Life 5 Dimensions Youth; Bolt on/off: EQ-5D with bolt on/off dimensions; ASCOT: Adult Social Care Outcomes Toolkit; FACT-8D: Functional Assessment of Cancer Therapy Eight Dimension; HASMID: Health and Self-Management in Diabetes; ICECAP-SCM: ICECAP Supportive Care Measure; IQI: Infant health-related Quality of life Instrument; NDI: Neck Disability Index; QLU-C10D: European Organization for Research and Treatment of Cancer (EORTC) cancer utility measure instrument; QOL-ACC: Quality-of-Life Aged Care Consumers; QOLIBRI-OS: Quality of Life after Brain Injury overall scale; SF-6D V2: Short-Form Six-

Dimension Version 2; SOSGOQ-8D: Spine Oncology Study Group Outcomes Questionnaire-8 dimensions; WOOP: wellbeing of older people.
Note: Shah (2020) used EQ-5D-Y and EQ-5D-3L

Among all of the measures, most of them were focused on physical and mental health change. Some studies generated utility weights for caregiver QoL measures considering wellbeing, health, and needs. Adult capability wellbeing measure ICECAP-A[192], the palliative and supportive care supported patient evaluation measure ICECAP-SCM were found and classified as wellbeing measure following consensus 1. The Australian health and care preference comparison research used the wellbeing measure ASCOT and EQ-5D to measure the relative preference across dimensions for the two measures. HASMID, QOLIBRI-OS, and CarerQol-7D are classified as wellbeing measure because of consensus 2 (all of the three measures evaluated life pleasure). DRU-I was classified as wellbeing measure according to consensus 3. In conclusion, some measures covered a broader wellbeing change [8] and this review found that 14 out of 65 articles valued measures covering wellbeing.

3.3.4 Attributes and Choice sets

The majority of measures were described by five attributes with five levels, with a range of 5 to 13 attributes. Thirty-two studies included a duration attribute (n=29) or included a 'death' scenario (n=3) to collect relative preference for anchoring. The range of duration levels of all the included studies was 2 months to 15 years. 1, 2, 5 and 10 years were the most common duration levels as recommended by the QLU-C10D valuation studies[193, 194]. However, other condition-specific PBMs and SF-6D valuation studies took 1, 4, 7 and 10 for duration level design. This study did not find any study which combined the two duration level designs together. Three studies reported conducting qualitative interview for deciding duration levels, but evidence from published studies was the most common source of duration levels. Most studies (n=28) selected the n-1 number of duration levels compared with other attributes. The methods employed by researchers to determine duration levels in their studies were varied. Three studies opted to conduct additional interviews to gather insights on

duration levels directly from stakeholders. Another common approach was to draw upon evidence from prior research as a basis for setting these duration levels, as the SF-6D valuation studies did. Moreover, it was observed that most research projects tended to select a number of duration levels that matched the quantity of other attributes in their study, maintaining a consistency and orthogonality in the level of detail across different aspects of the research. However, Lim's[126] EQ-5D severity-stratified study used 12 duration levels (from 2-month to 9 years), though there was no explanation for the duration selection process.

Valuation studies can present choice tasks with either paired scenarios [177, 195] or triplet scenarios[196]. Pairs (n=65) was more commonly used than triplets (n=3). The additional scenario was either described as dead or worst health state of valued measure. The total number of choice sets ranged from 28 to 960, with a mean number of 269 (Table 3). Over 60% (n=34) of all studies presented less than 12 DCE tasks per respondent, and the number of tasks varied from 7 to 28 (including dominant task or consistency test task). The respondents were asked questions about selecting their "preferred" health state (n=32), the "Best/better" scenario among the options (n=10), or the scenario that they would like to pick (n=4).

3.3.5 Study design and presentation

Studies applied mathematical algorithms to eliminate the DCE scenario number and generate an efficient design or used random selection/orthogonality array to decrease the number of potential pairs. The reported efficient design approaches were optimizing D-efficiency with non-zero (n=26) and zero priors (n=1). Informative prior values could be applied in minimizing the D-error in efficient design, where the priors may come from a pilot study or extracted data from published articles (e.g., taking Dutch EQ-5D-5L values as the fixed prior values for 15 EQ-5D valuation studies). Studies with prior distribution information and value uncertainty iteratively extracted priors with a Bayesian method (n=25). As recommended by the EuroQol EQ-5D-5L international valuation protocol, Bayesian efficient design has been applied by a larger

proportion of DCE studies (38% versus 30% from 1999-2018 review[84]). A non-informative prior value was applied to design a small-scale pilot study, followed by design update with the pilot data values and distributions. Other design strategies included fractional factorial design (n=8), C-efficient design (n=1), full/fractional factorial design (n=1), and others (hand selection and self-adaptive) (n=2). Four studies applied mixed design strategy, including both D-efficient and suppressing unrealistic/severe health states using hand selection (Table 4). The innovative adaptive DCE method presents certain uncertainties, particularly regarding the efficiency of the adaptive DCE design and the optimal way to present choice sets.

Table 4: Design method used

Design type	Approach	Identified study
Efficient Design	Bayesian efficient design	25
	D-efficient (with fixed/zero prior)	26
	C-efficient (with fixed/zero prior)	2
Fractional factorial	Randomized design and orthogonal method	8
Factorial design	Full factorial/fractional factorial design (Including adaptive DCE)	2
Other	Others (hand selection and self-adaptive)	2
	In total	65

Note: study using both efficient design and exclude combinations of dimension levels that were considered highly implausible in practice has been classified as efficient design

A study design with all attributes varied in each choice set provided higher statistical efficiency with a given number of respondents, yet it simultaneously increased respondents' confusion, misunderstanding, and dropout rate [44]. 25 studies presented the choice set with strategies to reduce the cognitive burden and increase respondent participation. These strategies included attribute overlap and visually attractive choice set presentation. 22 studies introduced within-dimension overlap, and eighteen studies highlighted the dimensions that differed within a choice set using different colors (yellow or light grey). Three studies presented the attributes that differed within a choice set before fixed overlap levels, where the other studies (n=62) presented all attributes including those that were fixed and those that differed. It was

found that health states described by a larger number of attributes (attribute number larger than 9) applied some degree of attribute overlap within the pair of health states in a choice set.

Included studies involved a randomization process of choice sets or sample randomization to increase face validity. 33 studies applied a process of blocking choice sets, including Balanced Incomplete Block Design (BIBD) or partial block design, to guarantee a balanced severity level distribution. Twenty-nine studies randomized the choice sets into blocks without stratification. Respondents in 6 studies answered a fixed number of unduplicated DCE questions, regardless of the attribute level overlap. Some studies developed in-block randomization: 16 studies randomized the choice set order in a fixed DCE block, 12 studies randomized the scenario sequence (left-right randomization), and 20 studies arranged the measure attributes in a random order (randomized dimensions). Studies with non-randomized blocking employed a weighted correlation strategy to minimize the average correlation between the blocking columns and all other design columns (n=1), or block with half random selected and half balanced (n=3).

3.3.6 Statistical analysis

Table 5 summarized the number of articles using different utility models employed. The main effect linear utility function (n=19) and main effect interacted with duration (n=26) were the most frequently used model functions. The main effect model captured only single-parameter main effects without interactions or extra dummies, while the main effect interacted with duration model estimated attribute coefficients with duration interactions with an additional duration attribute coefficient[44]. Both model specification forms assume that there was no dimensional interactions between PBM attributes[67]. Some studies considered the interaction between non-duration attributes[188, 197] or included an extra dummy to capture the impact of extreme health states (n=3). Shafie et al[198] and two other studies used an eight-parameter non-linear constrained model, where the parameter representing level 5 and one

parameter for levels 2, 3, and 4 (L2, L3, L4) were included. A hybrid model function (n=13) used both DCE and cTTO data mentioned, where the majority of EuroQol measure valuation studies included this model.

Table 5 Utility function and data analysis function

Characteristics	**Approach**	**Identified study**
Model function	Main effect linear utility function[1]	19
	Main effect interacted with duration (with and without constant time assumption[2])	26
	Main effect with extra term (dead dummy or worst/N3 state)	3
	Hybrid model[3]	13
	Main effect with constrained model (eight-parameter)	3
	Personal value function[4]	1
Regression function[5]	Conditional logit model (random/fix effect, scale-adjusted	48
	Mixed logit/latent-class logit model (heterogeneity model)	33
	Likelihood function (TT0/BWS with DCE data)	24
	Scale-assessment models/Poolability	7
	ZBT model with power function	2
	Mean individual preference	1
	Total identified articles	65

Note: 1. Main effect linear utility function is the model function with only DCE data and consider no dimensional interaction or co-effect with duration attribute.

2. Research considered both interaction with duration and extra term is classified as main effect with extra term.

3. Hybrid model is the Main effect function with cTTO or BWS data.

4. The personal value function is a self-adjusted health state valuation function, where the social preference is the average of personal preference.

5. Research can use more than one regression function.

For the regression model, the conditional logit model (n=48) was the starting point of choice data analysis. A conditional logit model is consistent with the random utility theory and assumes no scale or preference heterogeneity[199]. On the other hand, thirty-three studies applied the mixed logit model (n=24) or latent-class model (n=9) to control for individual heterogeneity. 24 studies used a hybrid model, which jointly

modelled both DCE and TTO preference data using a likelihood function. 7 studies considered the possible heteroscedasticity issue with conditional logit model, and estimated the scale effect with scale-assessment models[200]. The Zermelo Bradley Terry model (ZBT model) with an unilinear time preference[201] model appeared twice (more details in the following chapter) and the mean individual preference model showed one time (Table 5). Studies evaluated model performance with logical judgements: if the 'worse level' has higher latent or anchored disutility value, then the item coefficient would be regarded as inconsistent. Our updated review found over 60% of all studies reported some degree of inconsistency with the conditional logit model. However, there was no significant inconsistency rate increase with the DCE valuation result, compared with other valuation methods where these were included[84, 175].

3.3.8 Anchoring

DCE values were estimated on a latent scale. However, to generate utility values on a QALY scale, latent coefficients should be anchored on the 0 (full health) to 1 (dead) QALY scale, which can be done using a variety of different methods [87]. Sixty-one studies anchored the latent coefficient by using: extra TTO data (n=24); VAS data(n=3); duration attribute for estimating relative preference with time (n=29); re-scaling method with or without additional data (n=5).

There are 36 papers published on DCE_{TTO} and 4 with the triplet design with duration and death, including a second-hand data analysis study. The majority of DCE_{TTO} studies employed SF-6D and EQ-5D-5L measures, which are short measures focusing on health. However, DCE with duration were less used with long measure due to the cognitive burden of long measures and less recommendation of those measures by HTA institutes[202].

3.3.9 Design similarity

The selection of an experimental design is a crucial phase, significantly influenced by various "famous" designs. In an effort to guide future research and the selection of methodologies, this review endeavored to delve deeper into the underlying reasons for

the similarities and differences observed in study designs, including choice set selection method, anchoring strategy and information presentation. Our research revealed that the study designs for EQ-5D valuation, SF-6Dv2 valuation, and the cancer-specific EORTC QLU-C10D significantly influenced the methodologies: increased the use of EQ-VT (version 1 and 2) and the Australia EORTC QLU-C10D valuation method[203]. A considerable number of value set generation studies incorporated duration attributes found in the EORTC QLU-C10D valuation guidelines[193], and composite TTO (cTTO) data in analysis recommended in EQ-VT valuation[119]. One advantage of using the DCE with duration design was that the DCE data can be anchored without extra cardinal data [44], and the hybrid model with cTTO data used additional TTO information to increase its accuracy. Common DCE with duration levels for the duration attribute of 1, 4, 7, or 10 years/ 1, 2, 5, or 10 years were from the SF-6D v2 and QLU-C10D valuation models[203, 204]. In addition, the increasing trend of using D-efficient DCE design with priors and considering no dimensional interaction was influenced by the methods used by EQ-5D-5L, SF-6Dv2 and EORTC-QLU-C10D three studies as well.

While common designs of key PBMs focusing on main effects and interactions with duration remained predominant, there has been a noticeable shift in some studies to accommodate interactions between health attributes. This approach represents a perspective in study design, recognizing the complexity of health-related decision-making and the potential for different health attributes to influence each other. An ISPOR report[205] shows that estimating interaction effects among measure attributes should rely on quantitative analysis instead of assuming that the interactions are not statistically significant. More detailed methodology research on the significance criteria of attribute interaction is potentially worth exploring in the DCE health state valuation literature. Included methodology studies found that if interactions are important, these should be accounted for in the experimental design. In conclusion, the results of this review suggest that the methodological consensus identified in the Mulhern's review might be influenced by the measure that is valued, rather than academic

agreement[84]. Whilst this reflects a policy-making demand for PBM tariffs, it is recommended that the method and selections of the levels of the duration are examined further using qualitative research instead of copying the levels used in a published study or just relying on the most common levels identified during a review.

3.4 DCE Design options for long measures

Valuing long measures such as EQ-HWB presents significant challenges. These measures typically encompass a wide range of attributes and dimensions, making it difficult to design studies that can effectively capture and quantify such multifaceted concepts [144]. The published valuation studies using longer measures can provide us with valuable insights on the feasible design and information presentation options with classic statistical design methods. This section will introduce five feasible study design strategies and the attached information presentation options for valuing long measures: the efficient design, the orthogonal design, the Fold-in Fold-out design (FIFO), the Pivot design and the adaptive conjoint analysis design (taking PAPRIKA as an example). In the subsequent part of this section, the term "study design" denotes both the choice set selection strategy and the information presentation strategy (examples introduced in the Chapter 5).

3.4.1 D-Efficient design

A D-efficient design streamlines the process in DCE by reducing the choices from a full factorial set to the fractional factorial choice sets. This approach is not random but a calculated method aimed at minimizing the asymptotic variance-covariance (AVC) matrix in parameter estimations [206]. Given that these values remain unknown until data analysis, the design compensates by using the most precise available estimates, known as 'priors'. The determination of these priors can follow four methodologies: setting them to zero, employing fixed non-zero values from a value set, deriving them from distributions of a value set with parameter estimation, or updating them with data from pilot studies. Each of these options has unique implications for the study's accuracy and applicability.

This design strategy can be implemented using software packages available on Stata (where priors are not set), R, and Ngene. Each of these platforms offers unique tools and functionalities to facilitate the execution of the design strategy, catering to different preferences and requirements of researchers in the field. A good example is the EQ-5D-5L and EQ-5D-Y valuation studies following the recommendations of the EuroQol EQ-VT v2 valuation protocol[198, 207].

3.4.2 Orthogonal design with generator

The efficient design approach in DCE aims to balance orthogonality and efficiency. However, in practice, the designs are often non-orthogonal due to efficiency consideration. With inaccurate prior information and a large number of attributes, simulation data suggests that efficient design can become the most 'inefficient' and ‘misleading’, as the selected choice sets violated the minimum standard deviation assumption and the most of choice context may provide little information on respondents preference[206, 210] On the other hand, the orthogonal design with a generator operates differently. This method seeks to find a fractional factorial design using an orthogonal array and a self-constructed generator[85, 99]. During the design stage, all attributes and levels are treated as dummy coded variables, ranging from 1 to the number of levels of a given attribute. The orthogonal array is then identified from a pre-existing list, and a generator is used to create the other options for each choice set.

Compared to efficient design, the generator approach does not require accurate prior values for error calculation, or the judgement to define a specification of error term. This characteristic gives it an advantage, especially when dealing with new measures and long measures where prior data might not be available or reliable. The orthogonality generates a theoretical efficiency of 100% under the ideal situation.

3.4.3 Fold-in Fold-out design

The FIFO approach, introduced by Goossens *et al*., represents an application of the Hierarchical Information Integration (HII) method on the stage of valuation[118]. HII

has been conceptualized by Louviere *et al*., HII assumes that respondents automatically categorize attributes and simplified decision-making processes in behavioral studies by grouping individual attributes into overarching constructs[211]. This method operates on the assumption that respondents categorize attributes naturally and assign a singular value to each grouped construct. In the health state valuation studies, the dimensions are served as ‘natural categorization’ of attributes. HII facilitates the valuation of relative preferences between these constructs through sub-experiments of the attributes in each category, and it assesses out-of-construct preferences via bridging experiments. To select the choice sets, researchers can use the normal strategies such as efficient design.

For any FIFO question, all of the attributes are presented, and it has various sub-questions valuing certain number of domains. Each question presents overlapped attributes with the same level information (Fold-in) and separately presents attributes with varied levels (Fold-out). However, a notable limitation of FIFO is the absence of a dedicated software package for designing these specific types of DCEs, nor are there any instructions about how to select FI and FO attributes if each item represents a different domain. This approach, while practical, may lead to inconsistencies due to the independent design of each sub-question and the random selection of choice tasks for combination.

3.4.4 Pivot design

Pivot design builds on the theory of partial profiling, which is a strategy that addresses the challenge of dealing with a large number of attributes in studies, as discussed by Witt *et al*[212]. This approach presents only a subset of the complete attribute information, making it more manageable for respondents. In this setup, the 'pivot' refers to a set of attributes that remain fixed across both scenarios in a choice set, which presented above the varied attributes as background information. Respondents needs to complete a series of partial paired tasks with different pivots to generate a value set for all of the attributes[83]. One example has been that Craig *et al*., implemented the

pivot design in evaluating the 29 attributes of the PROMIS-29 measure[136]. This approach assumes that the respondent makes decisions based on the varied items, rather than the entire set[213]. An earlier application of this method indicated that using a partial DCE profile could reduce the variation in regression weights but there was limited publication with this method in the future[114, 214].

The pivot design shares similarities with the 'folded-in' approach in efficient design. However, there's a key difference: while FIFO design presents fixed attributes in a grouped format alongside separately presented changing attributes, the pivot design exclusively presents the two changing attributes without FI information. This distinction highlights the expectation to balance respondent burden and data accuracy, and one issue around valuing long measures is the optimal number of questions per respondent.

3.4.5 Adaptive conjoint analysis design

Adaptive conjoint analysis design offers innovative approaches to managing the complexity of choice tasks in studies. These strategies, including Rank Inclusion in Criterion Hierarchies with Extended Rankings (RICHER), Potentially All Pairwise RanKings of all possible Alternatives (PAPRIKA), and the less commonly used classical ZAPROS method, adapt to the respondent's selections and utilize transitive relations to refine the choice tasks. A key aspect of these methods is that the number of choice tasks is not fixed for each respondent and can increase if the respondent makes inconsistent choices[215].

The RICHER model focuses on modelling preferences using incomplete information derived from partial statements[216, 217]. Respondent are expected to make incomplete ordinal statements during choice tasks with less information scenarios. The scenarios are then enriched with additional information, allowing for a more nuanced understanding of preferences even with partial data. The RICHER method does not employ transitivity to reduce the choice set pool[215]. This could result in a higher number of choice tasks presented to the respondent, potentially increasing the cognitive burden.

ZAPROS (an acronym from Russian language, meaning ‘closed procedure near references situations’) has been proposed by Larichev and Moshkovich[218]. The original idea is to compare scenarios (or vectors) that have two values altered, positioned near a reference point. The reference point is a hypothetical state considered optimal across all dimensions. This process gradually finds the position of all of the health states on the line of utility value. A key challenge in decision-making processes, as noted by Larichev, is the potential for respondent inconsistency, which can violate the transitivity characteristic typical of discrete choice answers[218]. ZAPROS addresses this issue by eliminating dominated choice sets based on the principle of information transitivity. In essence, ZAPROS ensures that the preferences expressed are coherent and logically consistent.

Compared with non-adaptive designs, the PAPRIKA method is distinctive in its ability to generate both a personal and a social value set, while automatically accounting for heterogeneity during the adaptive choice set presentation process[219]. The PAPRIKA strategy starts by identifying all the undominated pairs and followed by valuing scenarios, consists of at least two degrees (attributes). This approach leverages the transitivity property of additive value models, which allows the survey software to identify all pairs that are implicitly ranked. The valuation process continues until all pairs have been assessed, either explicitly or implicitly. A separate binary search that uses individuals’ DCE results from the valuation of attributes to locate states better or worse than being dead. The notable and sole application of the PAPRIKA method in health state valuation is the creation of the New Zealand EQ-5D-5L value set[219]. This experiment involved an average of five binary search questions to determine the dividing line between states considered better or worse than death, and 20 adaptive DCE pairs to generate the value sets. Three levels (level 1, 3, and 5) from the EQ-5D-5L measure were used to construct the adaptive choice sets. The two median levels were determined through interpolation.

However, a significant challenge arises, when this method is applied to cases with more than 10 attributes, that the minimum choice set number is 126 (with second-order

and third-order stability questions), potentially rendering the method impractical for use for longer measures. This high number of choice sets can lead to respondent fatigue and reduced data quality, as participants may struggle with the cognitive load and time commitment required.

The ZAPROS and RICHER methods have primarily been applied to assess relative preferences, while the PAPRIKA method has seen broader application in stated preference elicitation and health state valuation. While ZAPROS and the PAPRIKA method are theoretically similar in their approach to handling choice sets and preferences, a key difference lies in their starting points. PAPRIKA does not begin with a fixed reference point, unlike ZAPROS. This distinction highlights the varied methodologies in adaptive conjoint analysis, each tailored to specific types of data collection and analysis requirements, particularly in the context of preference elicitation and decision-making studies.

3.4.6 Information presentation

Even with the most meticulously designed study strategies, health state valuation involving varied attributes presents a significant challenge for respondent understanding. To mitigate this and enhance respondent engagement, various studies have introduced more respondent-friendly presentation formats, including highlighting changed dimensions with attractive color, visualizing measurement results in the form of colored balloons, simultaneously presenting level descriptions with numbers and simplified question wording.

The literature reviews suggested that eight distinct highlighting formats were identified. These formats included dark shading for more severe levels, "traffic light" color-coding for different levels, grey/yellow highlighting for overlapped domains, grey/yellow highlighting for changed domains, bold/yellow level highlighting, alternative highlighting, and no color-coding. Mulhern *et al.*, evaluated a subset of these strategies (alternative highlighting, grey/yellow highlighting for changed domains, and bold/yellow level highlighting) in an online survey [208], suggesting yellow highlight was the most

efficient. Jonker *et al*., explored how level overlap and color-coding affect respondent dropout rates[209], suggesting a reduced the dropout rate by 4%. When applying this method to longer measures, two primary issues emerge: determining the optimal number of overlapped attributes and setting the appropriate prior values. These factors are critical in ensuring the effectiveness and reliability of the design, especially in more complex or extensive research settings.

Despite these efforts to make DCE tasks more respondent-friendly, a recurring limitation is the perceived insignificance of intermediate or adjacent attribute levels. Participants often report difficulty distinguishing between middle or intermediate levels, such as level four and five in the EQ-5D[61]. This issue of 'level insensitivity', with 16 out of 42 studies in the updated review reporting challenges in capturing sensitivities to these intermediate levels, may be difficult to combat with presentation optimization strategies alone.

3.4.7 blocking

Blocking refers to the selection of choice sets from the design that each participant answers in the DCE survey. This process can be either completely random (random blocking) or more sophisticated allocations. From the literature and ISPOR guidelines, six distinct blocking strategies were identified, offering varied approaches to grouping and presenting choice sets in surveys[123, 205]:

Random blocking: selecting the choice set for each block without duplication or generates a random number after efficient design.

Blocking with extra orthogonal array: using one extra column from an orthogonal array for blocking strategies[220].

Efficiency maximization by minimizing the correlation between the blocking columns: this strategy is often used in software like Ngene in the modified Fedorov algorithm. It focuses on minimizing the correlation between columns used for blocking, aiming to maximize the design's efficiency[135].

Efficiency balance by using a weighted D-error: this approach involves assigning weights to the efficiency of the overall design and the D-efficiencies of blocked sub-designs (typically, one-third to the overall design and two-thirds to the sub-designs). This is done after the initial pair selection, striving for a balance between overall design efficiency and that within blocks.[221].

Balanced incomplete block design (BIBD): BIBD is a classic design method. The "incomplete" characteristic indicates that not all of the selected choice sets presented in the block. As a result, the design method is recommended for the situation that there is a large number of attribute levels[222]. The BIBD with HWB measure keeps the information balanced in each of the block by allowing each attribute to vary together with another attribute the same time in each block. It uses algorithms, with *t* level attributes and *S* blocks of size *m*, to produce blocks with same frequency on *t*. as well as *t*'s combination, in each *m* among *S*. For any BIBD, the number of blocks each pair of attributes appears together is m(t-1)/(S-1).

Sawtooth Software level balance and near-orthogonal design: this strategy is specific to Sawtooth Software, ensuring each respondent receives a different block[205].

In addition to the prevalent blocking strategies, there are less frequently used methods:

The blocked fractional factorial design (BFFD), is defined by a specific blocking variable, which can be defined by the blocking generator[223]. Despite its potential applicability, this method has not been adopted in health state valuation studies, which limited its consideration in the broader discussion. Software tools may label different names for BFFD. For example, Stata (MP16) randomly generated a blocking variable, where a series of Pearson chi-squared tests are used to calculate the correlation with design variables, to produce the blocking with lowest 'association'. To simplify the descriptions, the name BFFD represents all these kinds of methods.

While evidence to ranking the superiority of one blocking method over another were absent, the literature review evidence provided us with useful information. Nearly one-third of these studies (22 in total, including those employing the EQ-VT approach)

opted for a balanced approach with regards to level distribution, such as the Balanced Incomplete Block Design (BIBD). However, the 'balance' blocking never represented perfect evenly distribution of attribute levels. When BIBD was applied within the confines of a specific efficient design framework, like D-optimality, it merely strived to emulate the balance inherent to a BIBD as closely as possible rather than achieving it perfectly[224]. Moreover, the strategy of minimizing correlation/association was employed in 6 studies. It was important to note that when non-random blocking applied, maintaining a balanced level was often prioritized with efficient and generator designs.

3.5 DISCUSSION

3.5.1 Summary of review

This review generates a richer picture of valuing health and wellbeing using DCE and updated findings from the published literature. Compared with the published reviews covering the time periods of 1999-2018 [84] and 2007-2018[175], this review has identified a larger average number of published studies in the reviewed years, and for the first time included studies published in the Chinese language. The research concludes that not only a diverse range of DCE methods were used in the health state valuation studies, but a widening range of countries launched large-scale experiments to test the feasibility of this method and reached positive outcomes. This trend indicates that DCE is a valuable and feasible methodology for valuing health and wellbeing states, including less educated populations in developing countries[225].

3.5.2 Study design

The popularity of anchoring with duration, question wording and data analysis is driven by "standarized" international protocols or sample studies, increasing the comparability of results. On the other hand, it must be noted that standardization may be a double-edged sword. Firstly, researchers should consider pros and cons of all feasible options before valuing new measures, instead of picking the standard protocol, such as the DCE design applied on EQ-5D-5L, EORTC QLU-C10D and SF-6Dv2 measure valuations. Secondly, deciding study design details, such as duration attribute levels,

should consider social-demographic factors and participant background. Thirdly, there is still no gold standard for study design, especially for long measures.

Some of the methodology consensus reported by the Mulhern *et al*[84] review is reinforced in the last three years. Online DCE with the general population has been more frequently undertaken during the COVID-19 pandemic. Online DCE is a less costly and more flexible option for a large-scale survey. However, it is worth noting that requirement to undertake surveys online means that participants require internet connection, an appropriate device and some level of computer literacy, and this may affect the representativeness of the sample completing an online survey, and data quality can be lower[226].

The reporting of models accounting for heterogeneity and heteroskedasticity becomes more common in recent DCE valuation studies, while the conditional logit model is still considered and compared with the results of models accounting for heterogeneity. Conditional logit model is advantageous because it utilizes all of the information in the regression, and heterogeneity models (i.e., mixed logit model) accounts for the demographic information collected in the valuation studies[227]. Heterogeneity models tend to be more promising practices with prior research group knowledge and large samples. On the other hand, Doherty *et al* and Wang *et al* evaluated attribute non-attendance[228, 229] with a conclusion that some respondents were less likely to consider the physical dimensions[230], which was a systematic bias violating the discrete choice assumption that the individual considers all the information and may not be identified under homogeneity assumption[231]. This heterogeneity supported the assumption that decision strategy was the main factor deciding the preference heterogeneity observed, instead of the demographic factors. With the stratified-group evidence or identified decision strategy pattern, preference heterogeneity should be considered from multiple perspectives.

3.5.3 Measure and selecting priors

The preference for study design with informative priors, with fixed and Bayesian priors,

increased. The review indicated that using non-informative for the pilot design and updated with Bayesian method is commonly applied. It was accepted that using Bayesian design could maximize the efficiency of the determinant[232], with a price of extra effort on prior data collection. However, the advantage may be exaggerated by its charming theoretical efficiency instead of real-world improvement. using non-informative priors and informative priors may not cause a systematic difference. Kesselsa *et al*[233] has presented with a case study that noninformative prior efficient design did not cause result variation with a sample size greater than 1000. A risk with small-sample soft-launch survey to get the prior information is the appropriate sample size to acquire sufficient variation[234]. It is reasonable for valuation studies with a large sample size to use non-informative priors, without pilot information update, in the design stage.

3.5.4 Remaining questions and limitations

Although there was a wide range of health state valuation applications using DCE, some remining questions from the previous review remained[84]. The first was around the modelling function used. The majority of DCE data was modelled using the main effect approach or the main effect interacted with duration approach, where the non-duration interaction term was not considered[202]. By using DCE design with duration attribute, an implicit assumption was the attribute interaction should be “interacted” with time as well[235], and a “zero time” condition is equal to the state of death. With few exceptions[236], attribute interaction without time was largely ignored. Besides, all of the respondents are expected to “choose” at least one health state with duration to be worse than death, as respondents refused this assumption, namely making their choice only based on duration, are regarded as using heuristic decision strategy. Norman *et al* discussed the influence of relaxing the “zero time” assumption by incorporating a dead state option in DCE_{TTO} design{Norman, 2016 #381}. However, this assumption has not been fully relaxed in the new design where a varied duration still kept. Chapter 6 will discuss other options in considering the interaction effect.

Secondly, there was no comprehensive feasibility and efficiency comparisons of various DCE study design strategies, nor any study known whether a paired design or triplet design is more appropriate. A triplet comparison increased difficulty and dropout rate[175] but provided more information with a fixed number of choice sets. Regarding the anchoring strategies, the DCE with duration design is straightforward, but may encourage respondents to answer the paired comparison questions with simplifying heuristics[42], and there is still no consensus on duration level selection. At this stage, this study cannot answer the question 'What is the most appropriate and feasible way of designing a DCE valuation of EQ-HWB valuation'. Further evidence is required to generate more integrated criteria, considering understanding, statistical consideration, response rates and cost efficiency for study design selection.

Thirdly, although there is some empirical evidence suggesting [230] that overlapping and color-coding design strategies reduced the dropout rate of EQ-5D-5L[237], there is no evidence with long measures. A larger number of questions per respondent increases the information collected but with a price of higher dropout rate and fatigue effect. Future qualitative and quantitative studies are required to help researchers make the trade-offs.

3.6 CONCLUSION

This review provides up-to-date information of health state valuation studies using the DCE method. The number of published studies continues to grow dramatically and there is more homogeneity in the methods used in the published articles, but this is likely impacted by the use of international protocols for some measures. The review intends to answer a key question particularly useful for this PhD research: is there any DCE designs that are potentially feasible for measures with 15 or more health and wellbeing attributes? 4 methods emerged and their characteristics summarized. Like previous reviews, this study did not find a 'gold standard' or consensus in the DCE health state valuation study design strategy or universally accepted criteria to evaluate the validity of included design strategies. This updated review surprisingly found that

researchers introduced more sophisticated modelling strategies, more straightforward DCE designs and better administrative strategy to minimize the data analysis bias, cognitive burden and data quality issues raised by previous reviews, though most of the explorations are still at an initial stage. Further research, especially qualitative research to assess the impact of different methodologies is recommended to inform practice in health state valuation using DCE.

Chapter 4 Systematic Selection of DCE Attributes

This thesis introduced the EQ-HWB measures and outlined the critical steps for designing a DCE survey in Chapter 2 and 3. The overall aim was to test DCE methods in valuing a health and wellbeing measure. However, the 25 EQ-HWB items have correlation and a DCE with that number of attributes would have high cognitive burden in the EQ-HWB DCE valuation study, leading to insignificant main effects. Therefore, selecting the proper attributes for the DCE valuation study is necessary to make the preference elicitation tasks feasible. In this chapter, I provided details on the steps for selecting attributes from the comprehensive EQ-HWB long measure to mitigate cognitive burden and avert collinearity in the DCE tasks. This follows the best-practice guideline of DCE design and presentation, where a large number of attributes is not recommended[238, 239].

All of the EQ-HWB attributes were evaluated with five criteria, by considering the evidence on dimensional completeness, item performance, stakeholder preference, international and cultural performance, and the qualitative consultation results.

4.1 Justification for item selection

The feasibility of this valuation approach is multifaceted. Firstly, it involves assessing the viability of using DCE for the EQ-HWB measure with a large number of attributes, as seen in recent examples like PROMIS-29[136] and ASCOT+EQ-5D[202]. This study defines “large number” as more than the EQ-HWB-S and acceptable for the respondents. Secondly, it explores the feasibility of valuing a measure that captures both health and wellbeing, a concept still in its infancy for wellbeing measures. Thirdly, it examines the reliability of generating health and wellbeing value set DCE preference data. The overall number of attributes should be reasonable for a modelling estimation, given the expected sample size.

Instead of indiscriminately increasing the number of attributes, which risks exceeding cognitive limits and yielding ineffective data, a focused feasibility study with a reasonable number of health and wellbeing attributes can offer more profound methodological insights on this stage. In other word, by selecting the attributes, it is ensured that the feasibility study result is solely related to the DCE method itself, instead of the measure attribute number. The second advantage is mathematical concise and comparable for the DCE design and data regressions[240]. An important consideration of statistical feasibility is the reliance of choice data regression analysis with a linear utility function on the assumption that attributes are independent and identically distributed (*i.i.d.*)[124]. With a large number of attributes with correlation, an assumption of multi-order correlations deviates from the standard DCE utility valuation function and can lead to misinterpretations of the results. The last advantage is the result generalization. The feasibility result can be generalized to the EQ-HWB-S valuation if we measured EQ-HWB-S attributes plus more core EQ-HWB attributes.

In conclusion, by selecting the EQ-HWB attributes, this study effectively concentrates on testing the design and statistical feasibilities. This strategic reduction in attributes ensures a more focused and manageable approach to this research objectives.

4.2 Item selection rules

The process of item selection commenced with the initial consideration of the nine EQ-HWB-S items, followed by a systematic evaluation of the remaining 16 items. Various criteria, including dimensionality, psychometric performance, item feasibility, stakeholder preferences, and cross-cultural performance, were assessed using the evidence generated in the E-QALY project development, refinement and testing of EQ-HWB and EQ-HWB-S[161]. The item selection criteria were outlined in Table 6. It was imperative to exercise caution during the item selection process to avoid inadvertently excluding core information[241]. A number of amendments have been made since the psychometric survey, including a revision of levels and the wording of control item. Given that the EQ-HWB is a relatively new measure and both of the instruments have

been given an experimental version status by the EuroQol, further revision of item wording may be possible but the domain structure should be stable. This section will rely on the psychometric evidence generated for developing the first version of EQ-HWB. As the importance of EQ-HWB-S attributes were supported by empirical evidence on the psychometric stage and international studies, all the items from the EQ-HWB-S will be included.

As introduced in the Chapter 2, E-QALY project collected data from various sample groups or published literature. The dimensionality was concluded by conducting literature review of QoL measures and interviews with patients, social care users, and informal carers about the QoL impacts. A test of item performance was conducted with the Patient and public involvement and engagement (PPIE) group in England. The overall validity and International & cultural feasibility tests were conducted with participants from UK, Argentina, Australia, China, Germany, and the USA, using both semi-structured interview and survey method. Finally, stakeholder and expert consultations decided the relative ranking or recommendation of each item for EQ-HWB and EQ-HWB-S[161]. The These evidence from E-QALY project was extracted, re-calculated and evaluated by this research. This section explained how the evidence was used and compared.

Table 6 Item selection criteria overview

Measurement Criterion	Criterion explanation	Rules explanation
Dimensionality	To include domains and attributes in the attributes selected for valuation based on verified evidence from literature review and the psychometric analysis. Consider for inclusion: - At least one item minimum per domain and a maximum of one item per sub-domain - All the EQ-HWB-S items - Use 'Bolt-on' evidence to identify items	Consider for inclusion: - at least one attribute for each high-level domain and maximum of one attribute for each sub-domain - The EQ-HWB-S items; - 'Bolt-on' dimensions for the EQ-5D measure[242].
Item performance	Item performance evaluates three primary characteristics of each item: item independence, item response distribution and overall validity performance. Consider for inclusion items with: - Low item correlation - Appropriate distribution ceiling and floor effect - High weighted validity score in each E-QALY research country	Consider for exclusion items with: - Spearman rank correlation matrix with over 0.7 correlation with items from other sub-domains; - high or very low proportion for the best or worst severity levels with cut-off points either >70% or <5%, except for the disease-specific item vision, hearing, coping, sleep, fatigue. - a weighted average score generated from the E-QALY data, with UK and Australia having a weight of 0.2 and 0.15 for rest of the countries. Score less than 3 would be flagged as less preferred for item selection and score less than 2 would be flagged as dropping.
Stakeholder preference	Stakeholder preference evaluated using the retrospective evidence from E-QALY consultations with summarized scores. Consider for inclusion items: - High consultation I vote - High consultation II ranking - Agreement from PPIE group Keep/Drop voting and cross validation	Consider for exclusion items with: - Consultation I voting result transformed into summarized scores (with 'Keep in' for 1, 'Drop' for -1 and 'Unsure' for 0). Higher score indicates a higher preference for inclusion. The evidence was considered with Consultation II result. Attributes with negative Consultation I score would be flagged - Consultation II interviews with stakeholders to generate evidence on selecting EQ-HWB-S items, reflecting the UK stakeholder preference after reviewing the psychometric evidence. The item would be flagged if the voting is less than 16. - PPIE views of draft items was summarized containing 4 PPIE sessions. PPIE group

		attitude for ‘Drop’ would be regarded as negative attitude while ‘Keep’ for positive. Negative PPIE attitude would be flagged.
International and cultural feasibility	This criterion evaluates internationally/culturally unacceptable by the international team. Consider for inclusion items: - Passed the CFA feasibility in each country	Consider for exclusion items with: - The confirmatory factor analysis (CFA) results in United Kingdom, Germany, China, Argentina, Australia and the United States, where attributes with fitness issue with certain domains and related attributes would be flagged.
Consultation and qualitative evidence	A summarization of meeting feedback, supervision team discussion and the focus group findings (Chapter 5) Consider for inclusion items: - General public sample can interpret the attribute information in DCE	Consider for exclusion items with: - qualitative consultation participants reported no understanding issue with the attribute, or consultation participants reported understanding issue but the supervisors believed it should not be excluded.

4.2.1 Dimensionality

The dimensionality assessed the completeness of EQ-HWB dimensional structure. Three rules were:

1. EQ-HWB-S attributes: *include all of the EQ-HWB-S attributes in the attributes selected for valuation using DCE***.** The nine classifier attributes were agreed to be important in the development of EQ-HWB-S, and hence this valuation study should include them. These items are Mobility, Daily activity, Control, Concentrating/ thinking clearly, Anxious, Sad/depressed, Loneliness, Fatigue and Pain.

2. Domain control: *incorporate at least one item for each high-level domain and a maximum of one item for each sub-domain in the attributes selected for valuation using DCE* to maintain the conceptual integrity of EQ-HWB and minimize the concept overlap. The high-level domains were Feelings and emotions, Activity, Self-identity, Autonomy, Relationships, Physical sensations, Cognition[161]. The dimensionality also considered factor analysis results (both confirmatory factor analysis and exploratory factor analysis)[161]. The evidence was considered with exploratory factor analysis and confirmatory factor analysis results (criterion 4).

3. Bolt-on attributes: “Bolt-on” of EQ-5D is the development of new attribute to “bolt-on” to the standard EQ-5D measure, to increase the sensitivity of EQ-5D and capture important aspects of health[243]. The significance of "Bolt-on" domains in health state valuation practice lied in their ability to address pertinent but uncovered aspects by EQ-5D, aligning closely with the practical concerns addressed by the E-QALY project. Bolt-on dimensions considered were those dimensions in the EQ-HWB that overlapped with the EQ-5D, but that were not in the EQ-HWB-S.

4.2.2 Item performance

The item performance considered the attribute independence and validity with future

respondents. The credible attributes and levels for DCE attributes must be relevant for distinguishing health and must be amenable to valuation[88]. All of the attributes included in DCE survey should be supported by evidence from potential users that the attribute captured HWB change (good performance) and had minimum level of collinearity as predictors in a regression model (high feasibility with DCE)[244]. There are three rules for attribute selection:

1. Correlation: EQ-HWB attributes with strong correlation score were highlighted. Multi-attribute outcome scoring adheres to von Neumann–Morgenstern (vNM) utility theory, relying on the first-order utility independence assumption[42]. The statistical analysis, implemented through the multinomial logit (MNL) model, assumes independence of irrelevant alternatives[46, 47]. Significant overlap between items selected as attributes in a DCE task poses challenges to the orthogonal design assumption and escalates cognitive burden, and also suggests that one item could potentially be considered redundant in terms of providing new information in the DCE task. Spearman Rank correlations were used, reflecting item-level correlations within and across dimensions, to identify high correlations (>0.7) attributes.
2. Ceiling and floor effects: attributes with high ceiling or floor effects were highlighted due to low universality. While it was anticipated that some disease-specific conditions, such as vision and hearing, naturally exhibit low prevalence and a skewed distribution, a more balanced distribution in other generic dimensions was expected. Balanced items indicated that general public had higher possibility to be familiar with various severity levels and understand the difference. Items with high ceiling and floor effects are limited in their ability to assess an improvement/deterioration for participants already at the ceiling (best level) or floor (worst level). On the other hand, generic items that had a very low proportion at the ceiling or floor may indicate the lack of validity of the item. The focus was primarily on skewness within the UK and Australia data, aligning with this current valuation study's geographical scope. Any generic

items exhibiting a skewed distribution were identified but not immediately dismissed, as most studies do not consider skewness as significant issues[245]. This distribution evidence would be weighed alongside other factors to determine whether to retain or discard flagged items. The assessment of skewness in psychometric tests would hinge on whether floor and ceiling effects reach or exceed 70%, or if the proportion presented below 5%. All decisions would be grounded in the results derived from psychometric analysis.

3. Overall validity: attributes with low general validity were highlighted. In the E-QALY study, each research team from the countries participating was tasked with evaluating the psychometric and face validity of each item and instructed to use a scoring system ranging from 1 to 4 for overall validity assessment with the evidence with equal weighting for each of the six countries[161]. Attributes with low overall validity score indicated a relatively poor performance in understanding and response [246]. However, considering that this current study will primarily be conducted in the UK and Australia, applying an equal-weight average score across all six countries is not appropriate. Therefore the original overall validity scores, derived from the psychometric analysis results, were revised to achieve a weighted average score. This approach assigns a higher weight to the scores from the UK and Australia (0.2), reflecting their greater significance in this current study. The remaining countries would each have a weight of 0.15. This method ensured a more representative overall psychometric validity score, aligning with the geographical focus of this valuation study. The attribute item wording could influence response patterns, comprehension[247] and factor rating[248]. Given that EQ-HWB attributes encompass both positive and negative wordings, this valuation study refrained from altering the attribute wordings. Instead, it investigated the impact of mixed wording during the qualitative consultation stage (Chapter 5), with a focus on identifying and potentially excluding attributes that posed challenges or were perceived as problematic by

participants.

4.2.3 Stakeholder preference

The valuation process aimed to incorporate items with high stakeholder preference, including decision-maker, academic, and public preferences. Two consultations in the E-QALY project yielded separate summarized evidence, with a total of 71 stakeholders participating both consultations, primarily from the UK (54% overall) and identifying as academics (65% overall)[246]. The first consultation with the advisory group informed domain and item selection (voted as keep, drop, unsure) for EQ-HWB. Voting results were generated after this initial consultation. The second consultation asked respondents to consider all evidence and recommend an item's inclusion level (strongly recommend, recommend, not sure, or do not recommend) for EQ-HWB-S. Consultation I provided a general score considering single-item inclusion appropriateness, while Consultation II forced stakeholders to make decisions after considering relative importance. Besides, one PPIE sessions were integral to the E-QALY project summarizing views on draft items categorized as 'Reject,' 'Undecided,' and 'Include.' This criterion considered the three sources of evidence[246].

1. Consultation I: a transformed voting results into summarized scores using the anchoring rule as 'Keep in' for 1, 'Drop' for -1, and 'Unsure' for 0. This study summed up the voting results and ranking scores to explore the key stakeholders' preference. The consultation I score attached same weight for all participants and an attribute with score lower than 50 were flagged.
2. Consultation II: an item with an above-average number of votes (n=25) would be strongly preferred for including in the survey, while attributes in the lower 25% (vote n<15) would be considered not including in the survey.
3. PPIE: an item with 'Reject' from PPIE perspective would be strongly considered for exclusion. This evidence was considered with the Consultation I.

An overall consensus would be achieved on whether items should be kept, revised or dropped, and whether important information has been excluded in the former stage.

This Criterion is not included in Table 3 as this is a retrospective check of selected items.

4.2.4 International and cultural feasibility

In the E-QALY study, exploratory factor analysis (EFA) and CFA were used to confirm the dimensions of the EQ-HWB[161]. The primary EFA and CFA model, rooted in UK survey data, underwent confirmation through CFA analyses across six countries including Australia[249]. The factor analysis evidence was used to identify specific attributes with fit concerns, where attributes with low factor analysis performance had higher risk of mis-specification of the hypothesized causal relations between latent factor and the stated preference data[250].

The CFA revealed the presence of 15 distinct factors. CFA data from the original E-QALY project was used to confirm item fitness with potential response model. Items fit with the original model were considered included. Notably, the CFA did not include all the EQ-HWB items, where coping, control, and usual activities did not exhibit loading onto any specific factor. All of the EFA and CFA data were from the E-QALY project and CFA evidence published or reported after 2022 (e.g., the replication of the original confirmatory factor analysis in China reported on 2024 EuroQol Academy Meeting) were not considered, as these updated evidence were re-confirmation with different sample[246].

4.2.5 Qualitative evidence

The selected attributes, evaluated based on the criteria mentioned above, subjected to discussions with the supervisory team (DR, CM, DS and RN), and underwent further discussion in the qualitative consultation (11 participants from Sheffield). Qualitative participants were asked if the attributes were understandable and distinguishable for each level (Chapter 5 reported the qualitative questions). The supervisors (DR, CM) participated in the qualitative consultations. Haode Wang and the supervision team discussed feedbacks. Face validity and feasibility for the future DCE, along with implications on DCE design strategies, were thoroughly considered during item

selection. Items deemed infeasible, ambiguous, or problematic in the consultation were excluded. A final decision was made about any attribute should be included or excluded after the discussion.

4.2.6 Data source

Data from the E-QALY project Phases I, III, IV, and V were considered. Integration of quantitative evidence[246] and recommendations from multiple individual consultations was undertaken to generate a comprehensive basis for item selection.

The Dimensionality criterion encompassed evidence derived from the domain structure established in the E-QALY project Phase I and included items from EQ-HWB-S. A literature review on EQ-5D bolt-on domains contributed insights into dimensions enhancing the validity of the descriptive system of EQ-5D [251].

The Item Performance criterion considered item-level correlation, supported by Phase IV item psychometric test outcomes. Ceiling and floor effects incorporated E-QALY Phase IV psychometric test distribution evidence for each item at the first and fifth (highest) levels, with overall validity scores transforming E-QALY descriptive evidence into numeric scores by international researchers in different countries[252].

The Stakeholder preference criterion considered insights from E-QALY Phase V. Quantitative results were derived from the descriptive outcome of Consultation I in each country for each item, concurrently evaluated with the qualitative PPIE result for making judgment when the two outcomes aligned. The International and cultural feasibility criterion drew conclusion based on E-QALY Phase III CFA results. Items with reported CFA model unfitness were highlighted as potential concerns.

The attribute selection was based on EQ-HWB UK version 1.0 (obtained October 2022).

4.3 Item selection result

4.3.1 Selection with criteria

Evidence from applying the criteria is reported in Table 7. The exclusion decisions have been made for several items in the item selection process: *Frustrated, Stigma/belonging, Unsafe, Discomfort severity, and Enjoyable activities* were excluded due to significant overlap (strong correlation) and/or unsatisfactory overall validity. The *Pain frequency* attribute was dropped due to domain overlap, while *Discomfort severity* was excluded due to sub-domain overlap and correlation with pain severity. The decision to exclude *Hopeless* was driven by evidence from Consultation II and correlation results. *Memory* performed less favourably in terms of correlation. The decision to exclude *Support* was influenced by indications from both overall validity and stakeholder preference, suggesting potential insufficient evidence for confident inclusion. *Coping, memory, and support* were excluded based on the consensus from Consultation I and PPIE, despite memory and support being mentioned in bolt-on reviews. *Self-worth* was excluded due to unacceptable results from CFA.

In conclusion, Frustrated, Stigma/belonging, Unsafe, Enjoyable activities, Coping, memory, support, hopeless, Self-worth were excluded. These decisions reflect a careful consideration of multiple criteria, including stakeholder input, psychometric analysis, and consultation outcomes, ensuring a comprehensive and informed item selection process with the available E-QALY evidence.

Table 7 Item performance

	Attribute	Dimensionality			Item performance			Stakeholder preference			International and cultural	Consultation and qualitative evidence
Rules		Domain	EQ-HWB-S	Bolt-on	Correlation	Ceiling and floor	Overall validity	Consultation I	Consultation II	PPIE	CFA feasibility	Expert and focus group participants
1	Vision	✓	×	✓	✓	N/A	O	✓	✓	O	✓	✓
2	Hearing	✓	×	✓	✓	N/A	O	✓	×	✓	✓	✓
3	Mobility	✓	✓	×	× Daily activity	× Floor effect for both	✓	✓	✓	✓	×	✓
4	Daily activity	O	✓	×	× Mobility	×	O	✓	✓	✓	✓	✓
5	Self-Care	✓	×	×	✓	× Floor effect for both	O	✓	✓	✓	✓	✓
6	Control	O	✓	×	× Loneliness, Support, Concentrating	O	O	×	✓	×	✓	✓

7	Coping	O	×	×	O	N/A	O	×	✓	×	×	✓
8	Memory	✓	×	✓	× concentration	O	✓	×	×	×	✓	✓
9	Concentrating/ thinking clearly	✓	✓	✓	✓	✓	✓	✓	O	✓	✓	✓
10	Anxious	✓	✓	×	× Loneliness, Support, Concentrating, control	✓	✓	✓	✓	✓	×	✓
11	Frustrated	✓	×	×	× anxious, control, sad	✓	✓	✓	×	✓	✓	✓
12	Sad/depressed	✓	✓	×	× cope, control,	✓	O	×	✓	✓	✓	✓
13	Hopeless	✓	×	×	× happy	✓	✓	✓	×	O	×	✓
14	Loneliness	✓	✓	×	× support	✓	✓	✓	✓	✓	✓	✓
15	Support	✓	×	✓	× Lonely, control, left-out	✓	✓	×	✓	×	✓	O
16	unsafe	✓	×	×	✓	× Floor effect for both and	O	×	O	✓	✓	✓

						ceiling effect						
17	Sleep	✓	×	✓	O	✓	N/A	✓	✓	✓	✓	✓
18	Fatigue	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
19	Stigma/belonging	✓	×	×	× Self-worth	✓	✓	✓	×	✓	✓	✓
20	Self-worth	✓	×	✓	× Stigma/belonging	O	O	✓	✓	×	×	✓
21	Enjoyable activities	✓	×	×	× Mobility, daily activity	O	✓	✓	✓	✓	✓	✓
22	Pain severity	×										✓
23	Pain frequency	✓	✓	×	✓	× Floor effect for both	✓	✓	✓	✓	✓	✓
24	Discomfort severity	✓	×	×	O	✓	✓	✓	O	✓	✓	×
25	Discomfort frequency	×	×	×	× Severity with pain	× Floor effect for both	✓	✓	O	✓	✓	×

Note: 1. × for negative evidence, not support including the attribute with the given evidence.

✓ for positive evidence, support including the attribute with the given evidence.

O for mixed or no clear evidence.

1. Dimensionality evidence derived from published literature, where O for the domain Criterion indicated a factor analysis issue and × means sub-domain

overlap.
2. Bolt-on evidence derived from the review article “A systematic review of the development and testing of additional dimensions for the EQ-5D descriptive system” and the bolt-on item selection research “Selecting Bolt-on Dimensions for the EQ-5D: Testing the Impact of Hearing, Sleep, Cognition, Energy, and Relationships on Preferences Using Pairwise Choices”. Other bolt-on articles have also been checked to make sure no items/domains neglected.
3. Item performance Criterion Correlation is from the psychometric test E-QALY item level correlations chart, the “How difficult was it for you to get around inside and outside (using any aids you usually use e.g., walking stick, frame or wheelchair)?” is combined evidence of inside and outside, Self-Care item (correlation 0.92 on average) is not flagged, Coping item shows mixed evidence (positive & negative item in correlation chart).
4. The ceiling and floor effect evidence derived from Table A.3: Summary performance of items for distribution and known group difference[161]. Only UK and Australia data evaluated. × indicates floor or ceiling effect in UK or/and Australia (“XX effect for both” means the effect identified in the two countries). The attribute is reported as O if all of the rest four countries showed floor/ceiling problem.
5. The Overall validity and Consultation I evidence derived from Appendix document E-QALY Item Selection Consultation. Consultation II results was from Appendix E-QALY Classification Item Selection Consultation Survey. More information about E-QALY design, process of Consultation I and II, and the data reported can be found in Chapter 2 or the overview paper[161]
6. The Stakeholder preference PPIE Criterion evidence was derived from E-QALY project patient involvement and result report “The role of patient and public involvement and engagement (PPIE) within the development”. Mixed evidence is generated for the domain of Self-Care, Hopeless (the old expression “nothing to look forward to”). Support and Hopeless got negative feedback.
7. If the consultation I and PPIE shared the same opinion on any item, then it should be strongly considered included or excluded.
8. If the consultation II shared similar opinion with the consultation I, then it should be strongly considered included or excluded. However, due to the item selection rule in consultation II, the single rejection from the consultation II data would be considered as suspicious and would be combined with correlation and bolt-on evidence to consider. The O indicated this item is on the edge (n=16) of voting out.
If the CFA feasibility evidence is unacceptable, then an item would be excluded, though this criterion is not applied to the EQ-HWB-S items.

It was decided to include fourteen HWB attributes on this stage: *Vision, Hearing, Mobility, Daily activities, Control, Anxiety, Depression, Loneliness, Pain severity, Concentrating/thinking clearly, Support, Sleep, Fatigue, and Discomfort frequency*. Nine attributes were from the EQ-HWB-S: *Mobility, Daily activities, Control, Concentration, Anxiety, Depression, Loneliness, Fatigue, and Pain severity*. The additional five items of *Vision, Hearing, Concentrating/thinking clearly, Sleep, and Fatigue* were included. However, the inclusion of Sleep was debatable as there was no evidence for exclusion, but mixed evidence observed in the Spearman correlation matrix. There was no compelling evidence to exclude Discomfort frequency, but the evidence was mixed.

4.3.2 Expert and supervisor team discussion

The selection of attributes was reviewed with the project's leading experts on meetings. During these discussions, concerns were raised about the 'Discomfort' attribute in the future DCE survey. The discussion noted that it might be overly broad, potentially encompassing a wide range of physical symptoms not covered by other physical items, as well as 'mild pain' symptoms, which would be more appropriately categorized under the 'pain' item. A re-evaluation of E-QALY qualitative evidence from healthcare service and care users supported this view, indicating that the 'discomfort' item might lead to ambiguities under the circumstances of quick decision[246]. Hence, it was concluded that 'Discomfort' might be unclearly defined and correlated with pain, leading to its exclusion.

The inclusion of 'Support' attribute was challenged by the experts. The item's wording, "I felt unsupported by people" combined with the response option "None of the time" created a double negative that could confuse participants if they read quickly, leading to misinterpretation of the best and worst states. However, it was noted that with clear presentation of item information, the proportion of misinterpretation remained low (See Chapter 5 for more information). 3 out of 11 participants realized that this was confusing while other participants understood the descriptions quickly. Therefore, 'Support' was

retained.

After evaluating all of the evidence, attributes *Vision, Hearing, Mobility, Daily activities, Control, Anxiety, Depression, Loneliness, Pain severity, Concentrating/thinking clearly, Support, Sleep and Fatigue* (Table 8) were included.

4.4 Conclusion

The item selection systematically evaluated E-QALY project evidence, qualitative consultation results and experts' opinion to evaluate each attribute by its dimensionality, psychometric performance, item feasibility, stakeholder preferences, and cross-cultural performance. 13 items met the selection criteria and passed qualitative consultations for the valuation study: *Vision, Hearing, Mobility, Daily activities, Control, Anxiety, Depression, Loneliness, Pain severity, Concentrating/thinking clearly, Support, Sleep and Fatigue* (Table 8).

Table 8 selected items

	Domain	Item	Levels				
			No difficulty	Slight difficulty	Some difficulty	A lot of difficulty	Unable
1	Vision	How difficult was it for you to see (using, for example, glasses or contact lenses if they are needed)?	□	□	□	□	□
2	Hearing	How difficult was it for you to hear (using hearing aids if you usually wear them)?	□	□	□	□	□
3	Mobility	How difficult was it for you to get around inside and outside (using any aids you usually use e.g., walking stick, frame or wheelchair)?	□	□	□	□	□
4	Daily activity	How difficult was it for you to do day-to-day activities (e.g., working, shopping, housework)?	□	□	□	□	□
			None of the time	Only occasionally	Some of the time	Often	Most or all of the time
5	Control	I felt I had no control over my day-to-day life (e.g., having the choice to do things or have things done for you as you like and when you want).	□	□	□	□	□
6	Concentrating/ thinking clearly	I had trouble concentrating/thinking clearly	□	□	□	□	□
7	Anxious	I felt anxious	□	□	□	□	□
8	Sad/depressed	I felt sad/depressed	□	□	□	□	□
9	Loneliness	I felt lonely	□	□	□	□	□
10	Support	I felt unsupported by people	□	□	□	□	□
11	Sleep	I had problems with my sleep	□	□	□	□	□
12	Fatigue	I felt exhausted	□	□	□	□	□
13	Pain	I had no physical pain in the last 7 days			□		
		I had mild physical pain in the last 7 days			□		
		I had moderate physical pain in the last 7 days			□		
		I had severe physical pain in the last 7 days			□		
		I had very severe physical pain in the last 7 days			□		

Chapter 5 Survey Qualitative Consultation

Qualitative data are recommended to understanding of how participants interpret health factors and consider health state preference elicitation tasks [166, 253]. Qualitative methods have been used to obtain attributes and condense the complex decision making strategies[235]. However, there has been no explicit investigation focusing on the DCE EQ-HWB valuation study design. A qualitative consultation was conduct, in order to provide deeper understanding of the interpretation of HWB information in the DCE context, and evaluate various designs and information presentation formats. 11 volunteers from Sheffield participated in the focus group. This chapter reported the research objectives, methods, and results.

'Design' in this Chapter refers to the specific information presentation format tested.

5.1 Objectives

The qualitative consultation topics were to explore the interpretation of HWB attributes, discuss the pros and cons of each DCE study design option, evaluate the feasibility of anchoring methods attached to each DCE design, and other nuanced information presentation topics.

- *Topic one: interpretation of health and wellbeing attribute information in the DCE context.*

The inclusion of HWB information introduced complexity to the valuation design process. This study aimed to engage the general public to test if they interpreted the HWB information as expected in the DCE choice set. Discussions around the concept of HWB attributes and information interpretation in the described health state. The discussion focuses mainly on the nuanced interpretation difference between the individual attribute information and its meaning in the DCE choice set.

- *Topic two: discussion of the pros and cons of various DCE study design strategies and information presentation methods.*

User-centred design principles were increasingly adopted in the development of DCE designs, advocating for the use of qualitative methods in revising designs. The characteristics of DCEs influenced survey response rates[254] and statistical efficiency[210]. This focus group consultation was to garner an understanding of general public views and preference on the feasible designs. The first two designs are partial comparison pivot design and FIFO design. Only two to three changed attributes in each “pivot” or “folded-out” section will be presented for the respondents to make their choice. This method decreased cognitive burden and make respondents more focused on all the revised information in choice making. The paired design with full information presented and triplet design with dead state are also named DCE_{TTO} and DCE with death, which are successfully used with EQ-5D, SF-6D and cancer QoL measures’ valuation{Wang, 2023 #612}. All of the four designs are supported by evidence found in literature review.

- *Topic three: investigation of decision strategies.*

Econometrics evidence supported the assumption that respondents adopted a range of “coping” strategies to reply the complex stated preference questions[255]. A rationality test question, mainly presented as dominated choice pairs, can identify the respondents who have logically inconsistent responses to the DCE questions in a DCE survey. However, this shed little light on how respondents replied to other undominated questions. It was important to explain understand the underlying decision-making strategies. This study discussed this topic about how respondents made their preference decision with the HWB information presented.

- *Topic four: examination of other design issues, including wording of warm-up questions, introduction and question instructions.*

The last effort was to gain insights to optimize the question wording and presentation of warm-up questions. DCE guidelines underscored the significance of qualitative evidence, emphasizing its irreplaceable role in informing and guiding unbiased design[256]. In the fourth

topic discussion, participants were asked about their design preference with the paired and triplet design warm-up questions, and necessary information in question instruction to prevent vague description.

5.2 Method

5.2.1 Focus group method

This research used a focus group method. This method recruited a group of general public to engage on specific topics relevant to the research question[257]. In contrast to one-to-one interviews, the focus group was an efficient qualitative method, allowing for the collection of qualitative data from more than one participant simultaneously[258]. This method was commonly employed to gain insights into attitudes regarding various research questions, i.e. attribute selection, and to elucidate explanations for certain behaviours[259-261].

5.2.2 Participants

The feasibility study tested the DCE valuation with the general public sample, which was the target sample of the future DCE survey. A consistent sample provided tailored responses to the specific question and directly relevant data, as recommended by *Coast et* al[253].

Participants were recruited via email using a volunteer list at the University of Sheffield. Participants from the volunteer list received an email advertising the study, with links to the information sheet and links to a short questionnaire where they could register with personal details and available time. Then the researcher (Haode Wang) selected and allocated participants to various focus groups, balancing age, education level, and gender in each group.

The number of focus group consultations adhered to the information saturation principle[262] Four focus groups were conducted, with 3-5 participants each time. The decision to include a fourth focus group was not pre-planned but due to the high no-show rate due to the University strike in 2023.

5.2.3 Procedure

Focus groups were conducted by the lead researcher, accompanied by one supervisor (DR

or CM). A topic guide (Appendix B) was designed with semi-structured questions covering aspects of EQ-HWB attributes, comparison of different presentation of DCE tasks, decision-making processes, time and death information interpretation, and fatigue effects with given numbers of DCE tasks. Semi-structured questions were crafted in a natural DCE information presentation order: attributes interpretation (after completing the EQ-HWB), attribute in DCE, decision-making strategy, and the information presentation.

Each topic consisted of two or three follow-up questions. Participants were required to complete the EQ-HWB (selected 13 attributes) at the start of the focus groups. After introducing the DCE and research aims, the focus group facilitator (Haode Wang) led the discussion and summarized the main points of each respondent after their round of speaking. Considering the fact that all respondents had no experience with DCE survey before, one warm-up question with a dominated pair was introduced to explain DCE tasks before topic discussion. The facilitator introduced reasons for discussion before initiating the discussion. The question asked in each focus group considered the information saturation that none of the questions discussed in all focus groups.

To facilitate a meaningful comparison, a mock DCE task was presented with the specific design. Respondents completed the mock design and shared their opinion about the design (see Appendix B for the order of presenting and topics discussed). The efficient design was the baseline design for practice question. Compared designs were:

a. Efficient design: efficient design has one label/attribute name followed by the levels of correspondent health states. Figure 4 shows a mock-up of an efficient design example. All of the attributes presented and overlapping attributes highlighted with yellow colour. Different approaches for presenting and highlighting differences were explored in the focus group (Figure 4).

Figure 4 Efficient design sample

	Life A	Life B
In seeing	No difficulty	Some difficulty
In hearing	No difficulty	No difficulty
In getting around inside and outside	Some difficulty	Slight difficulty
In doing day-to-day activities	Slight difficulty	Slight difficulty
You feel you have no control over your day-to-day life	Never	Sometimes
You have trouble concentrating/thinking clearly	Only occasionally	Only occasionally
You feel anxious	Sometimes	Often
You feel sad/depressed	Often	Only occasionally
You feel lonely	Only occasionally	Only occasionally
You feel unsupported by people	Only occasionally	Sometimes
You have problems with your sleep	Only occasionally	Only occasionally
You feel exhausted	Only occasionally	Never
You have physical pain	Moderate physical pain	Moderate physical pain
Which is better? Life A or B		

b. FIFO design: Unlike presenting the individual level information of each attribute, the FIFO design groups overlapping levels together[263]. In other words, only the levels of changed attributes can be 'folded-out'. The 'Folded-in' format facilitates respondents in quickly identifying overlapped attributes (Figure 5).

Figure 5 FIFO design sample

	Life A	Life B
Lonely Sleeping Having trouble concentrating/thinking clearly Daily activity	Slight difficulty/ only occasionally	
Mobility	Some difficulty	Slight difficulty
Anxious	Sometimes	Often
Sad/depressed	Often	Only occasionally
Control	Never	Sometimes
Unsupported by people	Only occasionally	Sometimes
Feeling exhausted	Only occasionally	Never
Vision	No difficulty	Some difficulty
Hearing	No difficulty	
Pain	Moderate physical pain	
Which is better? Life A or B	◌	◌

c. Pivot design: the pivot design relied on providing a subset of attributes, named 'Pivot', that were overlapped. Respondents make partial comparisons with the rest of the attributes[83]. For this study, the 13 attributes were separated into three subsets related to baseline physical activities (Seeing, Hearing, Getting around inside and outside, Doing day-to-day activities), Wellbeing and mental health goodness (Control,

Concentrating/thinking clearly, Anxious, Sad/Depressed, Lonely, Unsupported) and other physical health (Sleep, Exhausted, Pain severity). Respondents make separate choice with each pivot question (Figure 6).

Figure 6 Pivot design sample

Life A	Life B	Question 1
No difficulty seeing **No difficulty** hearing **Some difficulty** getting around inside and outside **Slight difficulty** doing day-to-day activities	**A lot of difficulty** seeing **No difficulty** hearing **Slight difficulty** getting around inside and outside **Slight difficulty** doing day-to-day activities	
◯	◯	**Which is better? Life A or B**

Life A	Life B	Question 2
Never feel you have no control over your day-to-day life **Only occasionally** have trouble concentrating/thinking clearly Feel anxious **sometimes** **Often** feel sad/depressed Feel lonely **only occasionally** Feel unsupported by people **only occasionally**	Sometimes feel no control over your day-to-day life **Only occasionally** have trouble concentrating/thinking clearly **Often** feel anxious Feel sad/depressed **only occasionally** Feel lonely **only occasionally** Feel unsupported by people **sometimes**	
◯	◯	**Which is better? Life A or B**

Life A	Life B	Question 3
Have problem with sleep **only occasionally** Feel exhausted **only occasionally** **Moderate** physical pain	Have problem with sleep **only occasionally** **Never** feel exhausted **Moderate** physical pain	
◯	◯	**Which is better? Life A or B**

d. Triplet comparison: in terms of the number of compared health states in each question, there are triplet and paired comparisons. A triplet design (with death) provides more information and opportunity to ask a better-or-worse-than-death question directly (Figure 7), akin to the classic Time-Trade Off (TTO) valuation method.

Figure 7 Initial triplet design sample

	Life A	Life B
In seeing	Unable	A lot of difficulty
In hearing	No difficulty	No difficulty
In getting around inside and outside	A lot of difficulty	Some difficulty
In doing day-to-day activities	A lot of difficulty	A lot of difficulty
You feel you have no control over your day-to-day life	Never	Sometimes
You have trouble concentrating/thinking clearly	Often	Often
You feel anxious	Most or all of the time	Often
You feel sad/depressed	Often	Only occasionally
You feel lonely	Often	Often
You feel unsupported by people	Only occasionally	Sometimes
You have problems with your sleep	Most or all of the time	Most or all of the time
You feel exhausted	Only occasionally	Never
You have physical pain	Severe physical pain	Severe physical pain
Better than being dead		
Worse than being dead		
Which is better? Life A or B		

Questions asked in each focus group were presented in Table 10. The attributes were presented with or without labels in the efficient design format, where label was the first column summarizing the dimension. Respondents were suggested to comment on attribute, level and the general design format separately. As the discussion included comparing with death, participants were informed and given the opportunity to leave the focus group discussions at any time.

5.2.4 Data analysis, coding and theme selection process

The focus groups were transcribed verbatim. All transcripts were analysed in *NVivo 12.*

Data was analysed using a thematic analysis framework, by familiarizing with the transcript, generating initial codes during checking the transcript quality, generating themes in the second reading, reviewing themes in the third reading, defining and naming themes after combining the overlapped themes, and drafting the final report[264]. The data were coded, and a thematic framework was established aligned with the aims of the study. This research employed a 'bottom-up' method to build themes, allowing the framework to evolve with the progress of data analysis. The codes were analysed and grouped thematically[265]. Anonymity was valued and pseudonyms (e.g., P1, P2 etc.) were used throughout the transcription.

5.2.5 Ethics approval

Ethics approval to conduct this research was granted by the University of Sheffield Population Health Ethics Panel (050569).

5.3 Results

5.3.1 The sample and question asked in each focus group

Twenty-one individuals expressed their interest after circulating the first round of recruitment email. Ten withdrew from the research due to time arrangement and UK education strikes in the early 2023. There were 11 participants who participated. The research was undertaken face-to-face at the Sheffield Centre for Health and Related Research (SCHARR), University of Sheffield. More participants details were:

- All the participants were from University of Sheffield (employee or affiliated institutes, none of them was student).
- 18% were male; 82% were female.
- 18% were high-school or below educated, 46% achieved an undergraduate degree or college degree, 36% had postgraduate degree.
- 18% were aged 20-30, 55% were aged 31-50, the rest responses came from participants of 51 or older.
- Respondents were enrolled in various field of working: postman, sports coach, equipment maintenance staff, research assistant, caregiver, language supporter and professor (Table 9).

Questions asked in each focus group were presented in Table 10.

Table 9 Participation information summary

		Focus group I (n=4)		Focus group II: (n=2)		Focus group III: (n=2)		Focus group IV (n=3)		In total*	
		N	%	N	%	N	%	N	%	N	%
Age	18 - 30	0	0.0%	1	50.0%	0	0.0%	1	33.3%	2	18.2%
	31 - 40	1	25.0%	0	0.0%	1	50.0%	0	0.0%	2	18.2%
	41 and above	3	75.0%	1	50.0%	1	50.0%	2	66.7%	7	63.6%
Gender	Male	1	25.0%	0	0.0%	0	0.0%	1	33.3%	2	18.2%
	Female	3	75.0%	2	100.0%	2	100.0%	2	66.7%	9	81.8%
Education level	Higher or secondary	2	50.0%	0	0.0%	0	0.0%	0	0.0%	2	18.2%
	College or university	1	25.0%	1	50.0%	2	100.0%	1	33.3%	5	45.5%
	Post-graduate degree	1	25.0%	1	50.0%	0	0.0%	2	66.7%	4	36.4%
Total number		4	100.0%	2	100.0%	2	100.0%	3	100.0%	11	100.0%

***In total column summarizes the characteristics of all participants.**

5.3.2 Thematic analysis

Four main themes were identified:

- Information interpretation: summarized how the respondents considered and understood the information.
- DCE decision making: summarized the decision-making pattern of respondents under various designs.
- Presentation of DCE questions (with sub-themes preferred design characteristics, less preferred characteristics and neutral): presented the likes and dislikes of design and information presentation formats.
- General suggestions

Subthemes were identified under each of these main themes (Figure 8).

Figure 8 Thematic framework for the questionnaire response

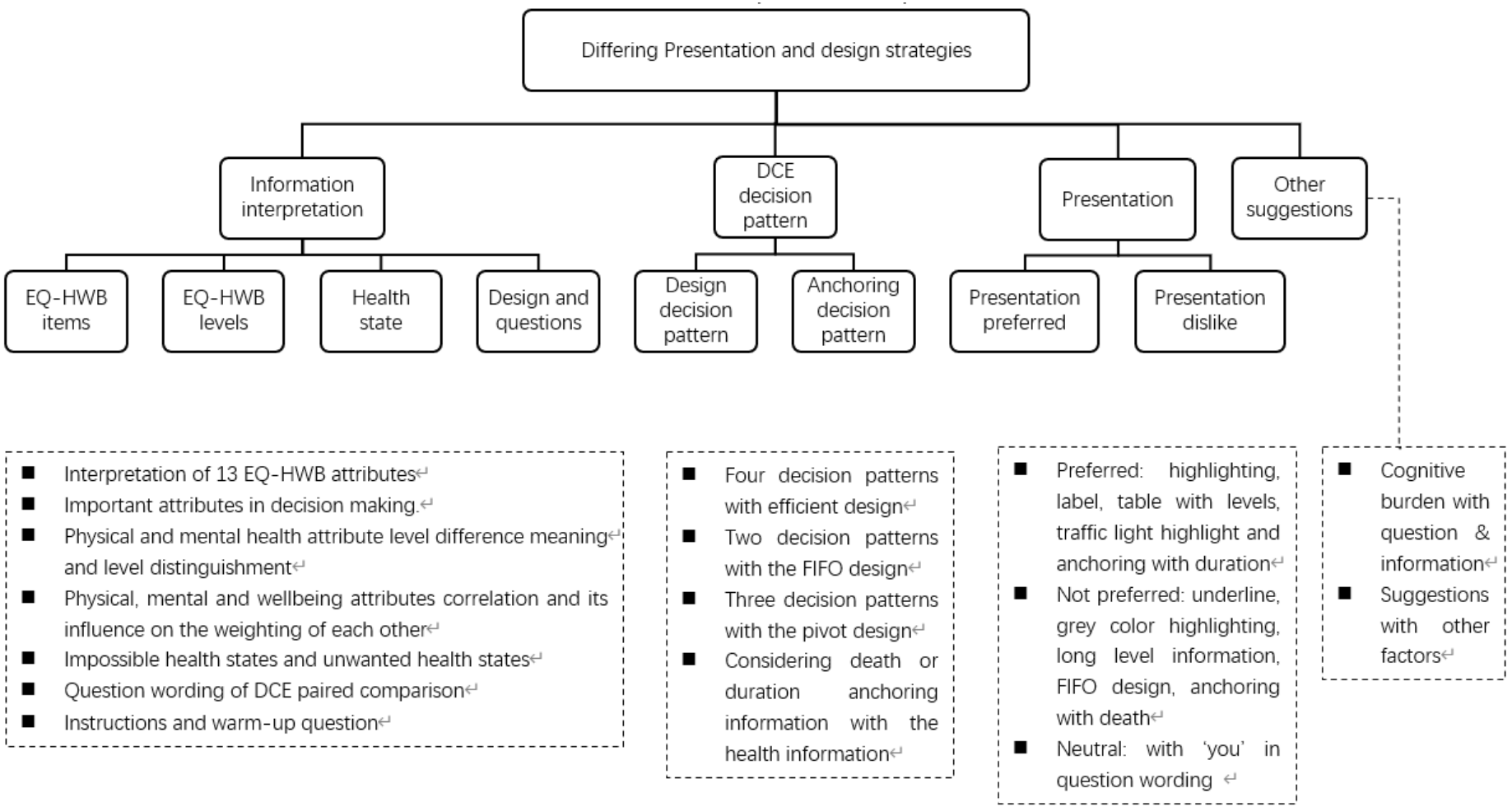


Table 10 Focus group questions

		Focus group I		Focus group II		Focus group III		Focus group IV	
		Y/N[1]	Changes[2]	Y/N[1]	Changes[2]	Y/N[1]	changes[2]	Y/N[1]	changes[2]
Focus group mainly discussed		General design		DCE wording		Anchoring and design		Design and wording	
Warm-up Question: the interpretation of EQ-HWB questions and DCE as a survey format	EQ-HWB attributes	Y		Y	Attribute order changed	Y		Y	Attribute order changed
	DCE warm-up question	Y		Y		Y		Y	In SurveyEngine Online format
	DCE instruction	N		N		N		Y	In SurveyEngine Online format
	How to ask DCE question	Y		Y	Ask the importance of subject	N	Ask whether have 'you' in question is important	Y	Ask where the DCE question should be placed
Question I: comparison of different DCE designs	Baseline presentation	Y	Ask attribute order, highlighting and decision process	Y	Health state more balanced	Y		Y	Ask where the DCE question should be placed
	Long information design	Y		Y	Ask whether double negative can be correctly interpreted	Y	Format optimized	Y	Order label question changed
	Labelled design[3]	Y		Y		Y	Format optimized	Y	
	FIFO design	Y	Use orange color for highlight	Y	Compare yellow and orange color information presentation	Y	Use yellow color only and health state levels balanced	N	
	Pivot design	Y		Y	Ask information saturation	Y	Order changed	N	

Question II: compare different anchoring	Anchoring with duration	Y		Y	Discussed more about their understanding with anchoring	Y	Time and health more balanced to prevent domination	Y	Ask where duration level should be placed
	Anchoring with death	Y		Y		Y	Discussed more about anchoring	Y	Ask death anchoring before presenting them the duration anchoring question
	Anchoring with duration and death	N		Y	More severe health states used	N		Y	
Question III: how many question and the less common health states	How many DCE questions to include	Y	Ask respondents whether 16 is an acceptable number	Y	Ask respondents whether 14 is an acceptable number	Y	Ask respondents whether 12 is enough for eliciting preference	Y	This question not asked
	less common health state	Y	Ask what health state can be defined as ‘worse than death’	N		Y	Ask what health state can be defined as ‘impossible health state’ (impossible combination of attributes)	Y	Ask a number of impossible health state samples from published literature
Question IV: others	Others	Y	Ask if understand what to do	Y		Y	Ask importance of introduction	Y	Ask if understand what to do

1. Y indicates this question is asked in the correspondent focus group.
2. ‘Changes’ indicates the semi-structured question change for certain topic compared to the latest focus group conducted.

There are five questions including the warm-up. The respondents were told on the discussion material, they were under warm-up or semi-structured questions and how many questions left.

3. ‘Label’ is the short first column before the two health state levels, to provide impression on what health problem it is.

5.3.3 Information interpretation

- *Attributes*

Participant distinguished the attributes presented as describing aspects of individual health and wellbeing. The number of attribute (13) was necessary to give an overview of health. Participants indicated a clear preference for the *Fatigue, Seeing, Hearing, Control, and Loneliness* in their decision-making process. The reasons for emphasizing *Seeing and Hearing* were related to the impact of vital physical function failure, hindering a large number of daily activities and affecting confidence. *Fatigue and Control* were seen as reflections of an individual's general status, where tiredness may lead to accidents and undermine productivity. *Loneliness* was designated as a valuable aspect of social wellbeing and a naturally concerning issue, even though the reasons extended beyond health.

Participants were uncertain about the definition of *Control* (control over work, which may be dictated by the supervisor, or just control of daily life), *Daily activity* (an ambiguous and broader concept), *Mobility* (getting around was be part of daily activities or only include travel outside the home), *Seeing* (whether myopia should be considered as a seeing issue), *Pain severity* (unable to justify if it was a state after using auxiliary means such as painkillers), *Depression* (hard to distinguish the levels and not clear if it was a diagnosed illness or contemporary mood) and *Support* (whether it was about real-life support or just their feelings about it).

Apart from the attribute wording, it was common that the attribute information interpretation was affected by personal experience, societal experience and imagination of the illness:

> Personal experience*: 'I think there are some things like occasionally feeling exhausted well that's what I do anyway.....(the existing situation) I did not know it was a health issue'* (*Focus group 1, P3*).
>
> Imagination*: '…..feeling anxious and sad and depressed if I could not see'* (*Focus group 1, P1*).
>
> Societal experience*: 'blind is a very high level disability…(will) affect the life very, very strongly'* (*Focus group 2, P6*).

The participants were questioned about their perceptions of the attribute ordering in the choice tasks and its impact on their choices. Grouping similar attributes (e.g. physical health, mental health, wellbeing) aided respondents in making connections. However, this approach led at least one participant feeling that they were being manipulated to focus on certain aspects or struggled with considering all attributes collectively:

> *Focus group 4, P11: 'I felt … being manipulated a little bit for me, think I cannot not think of other scenarios'*

- *Levels*

Respondents indicated interpretation issue with levels and showed a strong desire for considering severity with their personal situation or experience. Mild physical health issues (level 2 and sometimes level 3) were considered as acceptable and not that different to with perfect health, as these health conditions were seen as inevitable with age. The majority of the participants, who were retired or middle-aged employees, regarded the mild physical issue as being part of their life.

They also thought that distinguishing between the levels "occasionally" and "sometimes" was challenging. The levels at the severe end of the severity spectrum were confusing. As the health issue got more severe, it got 'different' to imagine. Besides, some attribute levels were not interpreted as expected. The description of the frequency of "loneliness" was interpreted as the frequency of having visitors or family gatherings. One respondent suggested using horizontal pain severity levels, instead of listing the pain severity as column, in the EQ-HWB measure.

- *Health states*

When evaluating a health stated described by a list of health problems with the given dimensions, participants automatically considered the sociodemographic of the subject and the uncertainty associated with the health issues. Respondents, in line with their varied opinions on attributes, unconsciously categorized health states, and the categorization influenced their preference decision:

Significantly bad health state: those with more than one extremely adverse attribute or severe issues with mental/physical health, such as solitary elderly individuals. Respondents tended to consider the significantly bad health states as invalid options in the choice questions. This could also be a feasible health state but '*not something that I want at all (Focus group 1, P4)*' in the preference decision.

Impossible health state: quote '*unrealistic stuff' (Focus group 2, P5).*' The health state was unpractical from their own perspective.

Normal health states: states respondents could imagine in decision-making. A choice activity between normal health states would consider and compare the attribute information of all options.

Respondents acknowledged a distinction between imagining health states predominantly in a 'hypothetical world' and recognizing the potential reversal of their answers in the 'real world.' They were conscious of these two perspectives and, although the 'hypothetical world' perspective was commonly adopted, it could change when considering health with duration. Respondents perceived the difference in their perspectives and demonstrated flexibility in switching between perspectives. For instance, participants mentioned

'*I would choose life B, but the reality is life A (with duration in the instruction, focus group III, P7)'*.

Mental health attributes had lack of attribute independence and were closely related with physical health. Having problems reported in mental health attributes may be a result of physical health and wellbeing. The weight of mental health attributes had higher correlation with physical health attributes, control, support and fatigue. In other word, the mental health issue was sometimes interpreted as a result of other attributes, expressing sentiments:

'*if you feel sad or depressed and maybe because, err, of lots of problems…having the additional problems of not being able to see, not being able to hear would make that even worse (Focus group 2, P6)*'.

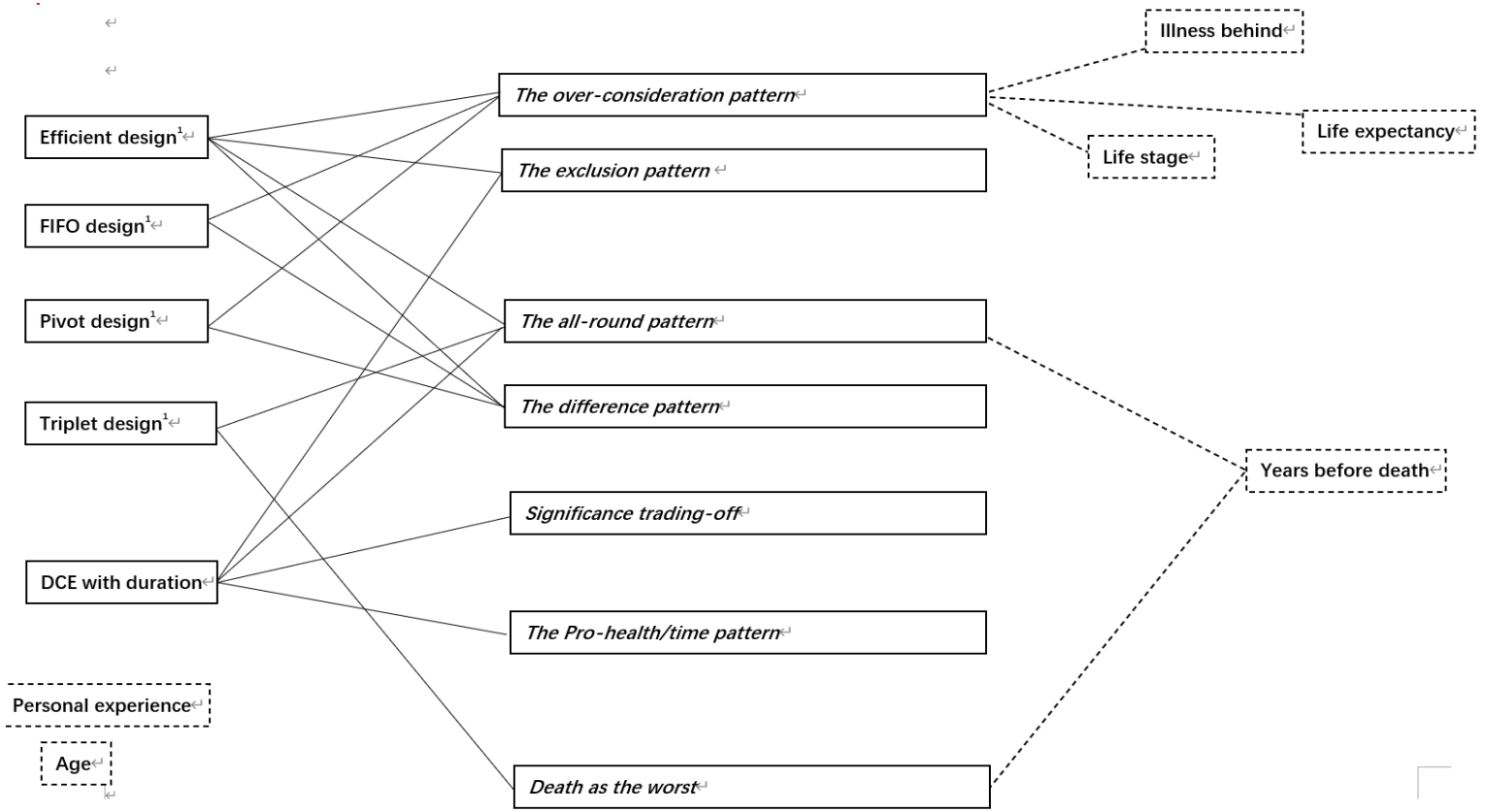
Mental health and wellbeing attributes' preference was also correlated with demographic

factors, especially age. However, as the introduction to the focus group embedded the concept of health, mental health and wellbeing to participants before the discussion started, participants were naturally tried to group the attributes into each category. They noted difficulty in splitting attributes into mental health and wellbeing categories, defining wellbeing attributes as part of mental health. When reading the labels, respondents tended to understand wellbeing as a kind of ‘feeling’ rather than a permanent state.

5.3.4 Decision patterns

The decision-making process among respondents revealed various patterns, and to avoid duplication, this research summarized decision patterns across different DCE designs. Figure 9 illustrated the decision patterns used with each design formats. It was important to note that an individual respondent may change their decision pattern, as they faced different DCE designs and questions, in the focus group.

Figure 9 Decision-making strategies



Note: 1. Efficient design: presents full information; FIFO design and pivot design: partial information design; Triplet design and DCE with duration: the efficient design with anchoring information

2. Continuous lines are connecting decision pattern and designs, while the dotted boxes and lines are other factors apart from the attributes

3. Decision patterns are merged if they shared similar characteristics in different designs. Partial information indicates respondents only consider aspects they value most instead of focusing on difference

Over-consideration pattern: the over-consideration pattern involved considering all attributes and making implicit imaginations about necessary background information not explicitly described. Imagination about individual characteristics, such as life stage, age, gender, and economic status were frequently reported. Participants made the health preference decision in a 'real' societal context, especially considering the background reasons putting themselves in the given situation. A common approach for making assumptions was respondent's own situation (e.g., experiencing the given imperfect health state at their age). Another approach was reasoning the context of given health state, i.e. why, how and the further influence of the health state, though this might lead to varied age assumption for the two scenarios presented in the DCE question. As one respondent explained below.

'….as you get further into these types of questions there's a lot more that you can, or you can consider behind it, so you have to make some assumptions…' （Focus group 3, P7）.

The all-round pattern: all-round pattern involved considering all of the information provided (but not additional factors). Respondents focused on the health description itself and took into account all of the information. The distinction between the all-round and 'exclusion' pattern (below) was that respondents in the all-round pattern iteratively evaluated all of the information after reading the highlighted attributes. They did not make assumptions about other social-demographic factors, as introduced in the over-consideration pattern.

'…how different they are, and then certain things will jump out, like you said, weighted so certain things would be like well that to me seems like a bigger difference so I focus it on' *（Focus group 3, P8）.*

For questions where the DCE task included a duration attribute, an all-round pattern respondent considered all information and making health and duration trade-off.

'it's about quality of life, and how long I've got.'（Focus group 4, P9）

The exclusion pattern: the exclusion pattern involved respondents having expectations for each attribute. If any attribute level could not satisfy the expectation, then the health state

would not be considered. In cases where two unwanted health states were presented, a random selection would be made.

'…. work out a balance between what you can "afford" to lose against other things isn't it now for me.' …'I did look at the deeper colours in the middle box, but then worked backwards from the deepest to the lightest to see in each side which I would be able to cope with'（Focus group 3, P7）.

'Narrowing down some like key things makes me think well actually.' … 'if you've got one particularly strong weighting (in one sub-question) against another one, so no difficulty seeing, a lot of difficulty, you know, it's gonna make the decision easy.'（Focus group 1, P4）

Difference pattern: the difference pattern involved making judgments by evaluating the varied attributes only. There were two ways of making decisions: the first way was difference counting, where respondents did not engage in an internal debating or trade-off process. Instead, they counted the better levels in each health state. A health state with more attributes rated as better than the other option would be selected. The second way was trading-off between varied attributes, respondents traded off the varied attributes with each other. The trade-off was limited to the attributes that differed between the two options. However, decisions based on the difference pattern could be unstable, as respondents always read less information, and may change the choice after a second reading.

'if the other ones are the same it's kind of like well they're exactly the same they're not gonna play much of a part in decision-making'（Focus group 3, P8）and

In relation to the fold-in fold-out design, where overlapped attributes were folded-in with different colours representing response level severity, one participant said:

'*I literally looked at the top box, it's a paler colour, but both lives have got the same issues… then jump straight to the middle (fold-out)'*（Focus group 2, P5）.

Significance trading-off pattern (with duration design only): the concept of 'significance' in decision-making reflected that respondents would only consider trade-offs time with health when any factor was significantly bad or good. There were two patterns within this concept: respondents would not consider the life expectancy only if the health state was deemed 'significantly bad'. Hence the life expectancy change could compensate the loss. Otherwise, they would continue to make decisions using one of the decision patterns mentioned earlier.

'I'd be looking through the information and then I kind of be thinking would this be significantly bad that it would not be worth me having the extra life.'（Focus group 3, P8）.

The second pattern within the 'significance' concept is the 'Time Difference Trading-off'. Respondents would consider HWB information only if the difference in duration levels between two states was significantly large, or the state with longer life expectancy was always chosen. When asked about the definition of 'significant' life expectancy difference, one respondent mentioned '10 years,' emphasizing that it could depend on the age and life stage. Additionally, a longer duration prompted respondents to consider factors such as how they would die and the impact on carers' feelings. Respondent reported willingness to live a shorter life under certain circumstances to leave a positive memory for loved ones (quote *'I want to leave them a good memory (Focus group 2, P5).'*).

'I mean, only if the difference is larger enough I would, (I1: trade it?) I would consider.'（Focus group 2, P5）.

- *Decision with anchoring information*

Respondents discussed the influence of duration information and death state on their comparison decision.

The better-or-worse-than-death decision was influenced by factors such as whether the health problem was manageable and whether there was enough positive reward to compensate for the negative aspects ('*although I can't see it, I know it's there, I can hear the birds (Focus group 1, P4).*'). Respondents considered the meaningfulness or hopelessness of a state when

making this decision. For example, a state that was regarded as meaningless or hopeless was generally considered worse than death. Instead of providing a description with the EQ-HWB attributes, respondents provided example with the specific disease or treatment that they were more familiar with, i.e. ventilator support, vegetative state, long-term depression, coma, and constant severe physical pain.

However, some respondents expressed that all the health states were better than immediate death. They emphasized the finality of death and the absence of a way to 'go back.' The ability self-adjustment and the value of hope were also factors that contributed to the perception that life, even in challenging health states, was preferable to death.

Another factor was the fixed duration information with the two health states. This research proposed “10-year in Health State A/B” and “4-year in Health State A/B" descriptions in focus group. Respondents considered the possibility of adaptation for longer duration but not for shorter duration.

“I think your main choice starts like that’s just for however long (imagine) you’ve got this is just the state and it’s not gonna change …… but I think in the real world my brain might evaluate it slightly differently because I think somethings I do think you might be able to alter for longer duration and other things you can’t. Whereas, for the point of the question (4 years), you’d have to be like well that’s just you’re like that…. (Focus group 2, P7) “

Respondents reported a decision strategy to maximize health rather than focusing on longer duration. They applied a heuristic decision strategy tailored for a health Discrete Choice Experiment (DCE) survey that only the health attributes considered.

‘Whereas, I would much rather have a shorter but healthier life’. （Focus group 3, P7）.

On the contrary, some respondents always prioritized living a longer life, choosing health states based solely on longer duration.

‘I think I’d still want to have those years even if they were not optimum health.’（Focus group 3, P8）.

- *Other factors*

Apart from the decision patterns, other factors influencing the decision were life stage and order of attributes. The evidence delved into the stage of occurrence of a health state on preference. Discussions directed participants to envisage living in a specified health condition for a certain number of years (X), without clarifying the starting point of these years. This ambiguity led some respondents to interpret the health state either retrospectively (X years before death, normally predicting health in older age with a general life expectancy), or prospectively (X years from their current age). This perspective could mirror the impact of age on health state valuation, where older age often correlates with a lower disutility value. For instance, severe health states might be perceived as 'normal' in advanced age, as one focus group participant noted

'*you kind of expect to have physical health problems when we get closer to, to the end-of-life expectancy (Focus group 4, P10)*.'

Despite the absence of a specific duration attribute in the triplet DCE mock task, health state duration influenced what constitutes a '*manageable health problem*'. The respondents concurred that many health issues might be bearable in the short term, for instance, over a year, but would become intolerable if endured for an extended period, such as the ten-year timeframe suggested in the survey. This was conflicted with the adaptation perspective.

The participants were asked about their perceptions of the attribute ordering in the choice tasks and its impact on their choices. Grouping similar attributes (e.g. physical health, mental health, wellbeing) aided respondents in making connections. However, this approach sometimes made them feel manipulated and struggle with considering all attributes collectively.

5.3.5 Information presentation preference

The Table 11 presented a design considering the qualitative suggestions. A preferred DCE information presentation format included characteristics that mentioned by at least one respondent from each focus group. Other suggestions, such as asking the choice question twice at the top and bottom of each choice set, was discarded out of the tidiness consideration and infeasibility with limited screen space. See Table 10 for other dislike and preferred presentation characteristics.

This section summarized the focus group consultation results on three core design questions: information presentation format (which design was preferred and why), attributes and levels presentation within each scenario (description of health issues and how to present the information), and choice set question wording and position (including both paired comparison and triplet comparison).

Yellow highlight

Centralized table design

Table 11: DCE design following preference

	Life A	Life B
In seeing	No difficulty	Some difficulty
In hearing	No difficulty	No difficulty
In your getting around inside and outside	Some difficulty	Slight difficulty
In your day-to-day activities	Slight difficulty	Slight difficulty
You feel you have no control over your day-to-day life	Never	Sometimes
You have trouble concentrating/thinking clearly	Only occasionally	Only occasionally
You feel anxious	Sometimes	Often
You feel sad/depressed	Often	Only occasionally
You feel lonely	Only occasionally	Only occasionally
You feel unsupported by people	Only occasionally	Sometimes
You have problems with your sleep	Only occasionally	Only occasionally
You feel exhausted	Only occasionally	Never
You have physical pain	Moderate physical pain	Moderate physical pain
You will live in the health state for	**4 years and then I die**	**5 years and then I die**
Which would you choose? Life A or Life B		

Labels used and with subjects in each attribute.

With order fixed and grouped together across similar domains

Duration attribute at the bottom

Suggested wording: Which would you choose?
Life A or Life B
Question at the bottom to help them know what to do with additional instructions at the top/outset for initial DCE tasks.

- *Presentation format*

The presentation format discussion focused on comparing partial profiles and full profile with overlapping.

When asked about the DCE design, every participant mentioned efficient design as "understandable." The potential advantage with partial design methods FIFO and pivot designs was considered as lower cognitive burden and evaluating the information in a more detailed way. This would enable higher participation rate. The negative comments were that partial profiles were misleading and strange, as the separate questions were difficult to evaluate as a single life:

'(the partial designs) ……. does not fit the purpose as well as the other one. The two designs are like three different incomplete lives, which is less like real life (Focus group 3, P8) .'

Another related theme was the duration and dead state information understanding. This discussion was presented with two efficient designs, one with duration attribute and the other with a third dead state. Some participants agreed that the DCE with duration was easier to understand and more comfortable.

'*If I'm given a timescale or just being dead, timescale would work for me (Focus group 2, P5).'*

Participants detailed the reasons of dislike such as religious belief and hard question with three states, and hardness to compare some possible state with the ultimate result dead. Other participants who supported the DCE with death stated that the duration brought more randomness:

"*…. I agree I think it's very difficult to answer. I mean, you literally could go and tick some in boxes and still not come up with an answer (Focus group 1, P4) ."*

There was no strong evidence support the exclusion of paired or triplet design. Both DCE with duration and design with dead state were practical and understandable with the 13 health and wellbeing attributes. However, a combined version with varied duration and triplet choice set was considered as too complicated.

- *Attribute and levels*

Participants discussed the preferred and dislike attribute and level information presentation. Design with attribute label, short levels, centre layout, varied attributes highlighted, and instructions at the bottom of the choice set, was preferred. Participants indicated that the short label was a respondent-friendly design to grab the key information. A centralized information layout, with cell borders, was suggested. Highlighting the varied part of information helped the *"Difference and important parts jump out (Focus group 1, P2)."*

The grey and traffic-lighting highlighting were not preferred. Participants supported presenting the question only one time at the bottom, some agreed that this made them know what to do fast:

"I think you're not gonna move on to the next page until you've ticked one of those boxes, so you know that it's something that you've got to do. …. otherwise, you just going down a list and maybe the last thing you read is moderate physical pain' (Focus group 4, P9)."

- *Question wording*

A full information presentation, including clear labels and a table format with highlighted differences, was favoured. Wording the question as second person pronoun, for example, "Your health" or "You prefer", captured individual preferences based on current information. Participants explained the reason:

'I feel less confident in making decision with for other people' (Focus group 2, P6)'

"Which is better" encouraged respondents to think more about health impact (Table 12). When asked any confusing wording with DCE questions and instructions, almost all participants supported adding instruction information that they could not return to previous questions. Participants should be made aware that the health state comparison was about preference rather than logical thinking. A statement saying "there was no right or wrong" would be helpful. Emphasizing important parts of the instruction and advising to have a pen ready for trade-offs were also mentioned, that respondents believed making notes were reading habits for some senior citizens.

Table 12 DCE design and presentation preference

	Factor	Reason for	Counts[2]	Reason against	Counts[2]	Outcome[4]
General Presentation[1]	Question number	■ **Maximum question number is 20**				√
	Information order	■ **Prefer fundamental health attributes first** *(Focus group 1, P2)* ■ **Prefer group together**: '*it makes it easier because you, you're thinking about similar things' (Focus group 4, P10)* ■ **Prefer fixed order: '***I would somehow argue though if you keep all of information order in a fixed pattern' (Focus group 4, P9)*		■ *'Not influence because it's not a long enough list to get tired of reading it' (Focus group 1, P3)* ■ **Not prefer group together**: *'I felt … being manipulated a little bit for me, think I cannot not think of other scenarios' (Focus group 4, P11)* ■ **Prefer changed order:** *'if you have a different order that actually make people read it more carefully and think about it more carefully' (Focus group 4, P9)*		√
	Survey instruction	■ **State the meaning of highlight: '***I mean it sounds obvious but when you sit down to do a survey you've got to take in so much information and you don't want to spend too long reasoning' (Focus group 4, P11)*				√
Information presentation	'Tell us which description you would prefer to live in'	■ **More confident in decision**: *'I feel less confident in making decision with for other people' (Focus group 3, P8)* ■ **Fulfill the respondent's expectation:** *'Align with what I expect to do' (Focus group 2, P4)*	6	■ **Other question wording options:** *'Suggest imagine someone will live in life X for n years with duration' (Focus group 2, P6)* ■ *'Which would you choose, life A or B is better' (Focus group 4, P9)?* ■ *'Saying ''think about your life and you had to live that life, which one would you prefer'' '(Focus group 3, P8).*	7	
	Question position	■ **Know what to do faster:** *'I think you're not*	3	■ **More obvious than the bottom:** *'obviously at*	2	

at bottom		*gonna move on to the next page until you've ticked one of those boxes, so you know that it's something that you've got to do. …. otherwise, you just going down a list and maybe the last thing you read is moderate physical pain' (Focus group 4, P9).*		*the top it really stands out' (Focus group 4, P10).*		
Short label used to describe the attribute	■ **Respondent-friendly design:** *'Easier to see things when the alignment was in the middle' (Focus group 1, P4).* ■ **Have subject in label***: 'it's not a massive difference whether it's you or I, both preferable to just the verb, cos that kind of disassociates' (Focus group 4, P9)*	5	■ **Some labels are confusing:** *'Just saw mobility and daily activity and control might not be 100% clear' (Focus group 1, P4).*	3	√
Long sentence used to describe the level in each attribute in each option[5]	■ **Understanding better:** *'makes more sense and gives more context' (Focus group 3, P8)* ■ *(For double negative) has the full sentence next to it makes it clear (Focus group 3, P7)*	3	■ **Too many words:** *'Difficult to read and adding potentially unnecessary layer' (Focus group 1, P2)*	3	
Table	■ **Reader-friendly design:** *'Clearer about which label goes with which thing' (Focus group 3, P8).* ■ **Centralize the information:** *'Have everything centralised in the line' (Focus group 3, P7).*	7	■ **Misleading: '***Look like go out to the categories' (Focus group 3, P8).*	1	√
Level bolding	■ *'Better than underline' (Focus group 2, P6).*	2	'(Underline) good if with table' *(Focus group 1, P3)*	1	√
(Yellow) Highlighting	■ **Grab the different parts fast:** *'Difference and important parts jump out' (Focus group 1, P2).* ■ **Grey is less preferred:** *'more difficult and having to concentrate' (Focus group 1, P2).*	7	■ **Yellow is less preferred:** *shouty in my face and hurts my eyes (Focus group 2, P5).*	1	√
Traffic light	■ **Severity distinguished:** *'the severity goes from pale to dark and easier to look at and decision easier '(Focus group 2, P5).*	3	■ **Unclear gradient of colour:** *'harder to distinguish the gradient of colour and not enough difference between the colours' (Focus group 2, P5).* ■ **Strange design:** *'colour palette like websites*	6	

				back in the day that not fully done' (Focus group 3, P8).		
	FIFO design	■ **Less information to read:** *'simple and easier' (Focus group 1, P1).*	2	■ **Confusing design:** *'It's confusing and those sections almost have no relevance' (Focus group 1, P3).* ■ **Misleading:** *'Does not fit the purpose as well as the other one' (Focus group 3, P8).*	7	
	Pivot design	■ **Consider information in detail:** *'Get down to detail on certain elements and separate the different parts' (Focus group 2, P5).* ■ **Lower cognitive burden:** *'less factors it makes it a lot quicker decision' (Focus group 3, P8).*	4	■ **More information considered overall:** *'not narrowing down what's important to me' (Focus group 1, P4).* ■ **Strange design:** *'like three different incomplete lives, which is less like real life' (Focus group 3, P8).*	7	
DCE design	Anchoring	■ **Easier to understand with duration:** *'If I'm given a timescale or just being dead, timescale would work for me' (Focus group 2, P5).* ■ *'(triplet) very hard question to answer and leads to random answer' (Focus group 3, P8).* ■ *'(triplet) unless in that situation, find it very hard to compare with death' (Focus group 1, P2).* ■ **More comfortable question with duration:** *'(triplet) uncomfortable judging whether they should be dead or alive, and it is very private' (Focus group 1, P4).* ■ *'(triplet) taking people on an uncomfortable journey of thinking about it. Want think positive '(Focus group 1, P2).* ■ *'(triplet) religious beliefs affects' (Focus group 1, P4).* ■ **More realistic question with duration:** *'(triplet) unrealistic, as you wouldn't compare the health state to death directly' (Focus group 2, P6).*	10	■ **Anchoring should not be included:** *'Don't really like considering death and life expectancy' (Focus group 1, P2)* ■ **No need for duration information with triplet design:** *'approaching it without having had all this background and help through' (Focus group 4, P10)* ■ **Prefer anchoring with both duration and dead state***: '…set one duration at the beginning rather than putting different durations in different questions.' (Focus group 4, P9)*	2	√

Duration at bottom	■ **Less confusing:** *'Less shock in the health question' (Focus group 3/4, P8/P11).* ■ **Health first satisfy the expectation:** *'Feel like where it should go' (Focus group 3, P8).* ■ *'Have chance to think health first' (general)* ■ *'Think about different life more' (Focus group 1, P4).*	5	■ **Consider time first better:** *'First is better for considering first'(general).*	2	√

1. The general presentation questions were asked only in focus group 4. No counting reported here.

2. The counts column presents how many times one certain theme mentioned and is a way of demonstrating how commonly this was raised.

3. Quotes of respondents are given with quote marks and the boldened words are paraphrases.

4. The 'outcome' column shows the outcome following review across the focus groups. A '√' means most of the participants (60% or above among all focus groups) agree and this format has been accepted for the final design.

5. The long level is to describe the health problem with a long sentence without a label for the first column. More information about the labels can be found in the Table 2 footnote. See Appendix B for the focus group topic guide.

5.3.6 Other suggestions

- *Number of Choice sets*

The discussion started with asking whether completed 12 tasks was feasible, and all of the participants agreed that 12-20 tasks could be finished in the 30 minutes. However, respondents also mentioned that the acceptable number and expected time depended on the decision strategy employed:

"I can take 14 questions, but not all, I, I looked at the majority of Life A to tick, to tick it. This one (pivot), I was imagine- I was able to mix and match. So I could take a part of it from here, a part from here, a part from here. When you look at the first one, it's, it's either set A or set B (Focus group 2, P5)."

- *Understanding and cognitive burden*

Thirteen HWB attributes were acceptable and reasonable for the respondents to complete the DCE questions. The first reason was that the DCE scenarios were considered to describe overall health, where no attribute was unexpected. The second reason was that the respondents already had an expectation of facing a large amount of information after reading the instructions:

"I think my thought would be that like if you're thinking about health, like someone's health and wellbeing overall, you are gonna have to ask quite a few questions and weigh them up and things in your mind because there's a lot of different aspects to go into it, so I feel like that's fine (Focus group 3, P8)."

5.4 Implications for study design

The focus group findings indicated that the respondents understood most attributes of the EQ-HWB and the DCE questions as expected. It provided evidence that participants could undertake a DCE survey to value the EQ-HWB 13 selected attributes, with an extra duration information or a third option dead. However, *Fatigue, Loneliness and Control* were less clear and the level 2/4 were less distinguishable. Results suggested that the FIFO and pivot design may not be an ideal design for valuing the health preference than efficient design. Anchoring

with duration or with dead health state were advocated.

A clear observation was that respondents lack of a consistent decision strategy to consider all the information provided when making their decisions. It was possible that using heuristic decision strategies and considering information unrelated with health interfered with evaluating the HWB information. It might lead to biased parameter estimation and the appearance of false preference heterogeneity, as noted in the literatures[255, 266-270]. A heuristics research identified seven common heuristics in health-related DCE[270], and the theme revealed the presence of five of these heuristics: the choice set format, lexicographic preference, dominant decision-making behaviour, attribute non-attendance and level non-attendance. This finding led to the suggestion that an appropriate data analysis method or follow-up question needed to collect more information about the decision strategies.

The preference for certain DCE design and information presentation formats was a recurring theme throughout the focus groups. Efficient design that incorporated overlapping and highlighting, evidenced by the second and third discussion topic, was deemed most suitable for valuing the preference of EQ-HWB 13 attribute. This approach also enhanced comparability with other studies. Respondents viewed the FIFO and pivot design as three partial questions to consider separately, which was conflict with the partial design assumption and increased the number of questions. The paired and triplet designs were kept for future DCE survey. The attributes information was given in a separate label column, while the level information was described as short severity/frequency of imperfections (Table 11). Despite the conclusion with design and presentation, it was important to note that there were very few instances in which participants expressed clear preference to one specific anchoring methods. Additionally, the perceived importance of life expectancy and the imagined social factors influenced respondents' decisions. The 'focusing effect' with time amplified the heuristic in DCE decision, while an uncertain starting point led to random decision-making for some participants [271]. To improve the understanding of DCE choice set, the question wording changed from “which you prefer” to “which would you choose”, making respondents confident with choosing from their own perspective.

In conclusion, a preferred design following the focus group evidence incorporated several key

features: full information presented with labels, centrally placed short levels, varied attributes that were highlighted and question instructions located at the bottom of the choice set. It was decided to not incorporate numbers reflecting the severity levels as this may cause confusion for some participants around how to interpret the numbers. Both paired design with extra duration attribute and a triplet design with dead state remained.

5.5 Limitations

The primary constraint for this qualitative work was small sample size, stemming from resource limitations and challenges in recruiting participants during national Universities strikes. Another limitation was the homogeneity of the participant pool: all participants were recruited from the University of Sheffield volunteer list. This raises concerns about the generalizability of the findings, especially that the future international online survey aimed to encompass both the UK and Australia, yet no Australian participants were involved in these focus groups. The possibility of desirability bias was also acknowledged, where participants might have provided responses they believed were desired or unproblematic, particularly in response to semi-structured questions[272]. In the group discussions, instances of disagreement fear and shifting opinions were observed, especially when respondents realized they were in the minority. Some respondents, i.e. the P8 from focus group 3, had a higher willingness to share their opinion. Their preference tended to have a higher weight in deciding the DCE format.

To mitigate these limitations, deliberate efforts were made during participant recruitment to include individuals with lower education levels and to balance socio-demographic characteristics within each group. The facilitator took steps to ensure a comfortable environment for participants and subtly encouraged those with dissenting views to articulate their reasoning. Follow-up questions were crafted based on participants' narratives, focusing on attitudes rather than cultural backgrounds or peer pressure, to elicit more authentic and diverse perspectives.

Chapter 6 Valuing Health and Wellbeing Using DCE: Study design

The overall aim in the development of this feasibility valuation study was to test the feasibility of valuing the long measure EQ-HWB using the DCE method, and to compare preference influence of designs and countries. Based on qualitative evidence (Chapter 5), this Chapter introduced what exactly design method the valuation study used, what presentation format tested, and how valid results got with the given sample size under the limited budget.

Secondly, extrapolating beyond the feasibility itself, this research has three core questions that provided insights with the data: (1) the influence of study design on HWB valuation results; (2) the influence of attribute order on preferences, and (3) whether the findings were comparable in UK and Australia. Additionally, this study tested the relative merits of several alternative model specifications, including the additive model, cross-attribute level effect (CALE) model and the rank-ordered logit model.

This chapter start with clarifying the research question. Then introduced the method used for study design, choice set selection, survey design, surveys implementation, and the data analysis. To distinguish from the wording '*study design*' used in Chapters 3 and 5, this Chapter uses '*design format*' to represent the choice task and anchoring design (i.e., DCE_{TTO} and DCE-Death) options and used '*design"* to represent the DCE design strategy for selecting the profiles (i.e., generator design and efficient design).

6.1 **Research questions**

The order of information: does the attribute ordering of health and wellbeing affect choices and answering times?

A potential methodological issue for valuing health state is the order of HWB dimensions presented to respondents. Dimensions can be presented in the standard order, consistent with the PBMs, or in a revised-order to prevent the influence of overall

perception of health states. Mixed evidence revealed on this topic with EQ-5D-5L measure[273, 274], and study reported no significant impact with QLU-C10D measure[275]. Compared with EQ-HWB, both EQ-5D and QLU-C10D describe health states with smaller number of dimensions. When using the WTP DCE design, a prediction is the attribute order impact the relative preference and tendency to engage with the given information[276]. The focus group participants reported remembering dimensions listed near the choice question better (See Chapter 5.3.3 for more information). This study compared preference elicited with DCE question in varied attribute order, to inform the study design and dimension randomization in future HWB valuation studies.

To explore this hypothesis, the study used DCE_{TTO} with two different versions of attribute ordering. The first version follows a 'standard order' that aligned with the order of EQ-HWB, with physical health attributes being listed first. Conversely, the second version adopted a reversed order, with wellbeing attributes listed first, followed by mental health and physical health attributes.

Duration and death anchoring comparison: do different design formats and anchoring methods leads to different results?

DCE with an duration attribute (DCE_{TTO}) allows valuation study to generate anchored value set with DCE choice data[44]. Since Bansback *et al* reported this method in 2012, the approach gained popularity in the EQ-5D-5L and SF-6Dv2 valuations[123]. A major uncertainty around the episodic random utility theory (episodic RUM) DCE_{TTO} grounded, as discussed by many researchers, was the linear time preference assumption[230]. A standalone solution in the literature was to optimize the anchored value by design a parameter accounting for the discounted “value of time”[230], or consider the latent utility of death with non-linear time preference assumption[129]. However, these methods relied on the assumption of respondent’s preference on duration, potentially leading to a new type of bias. On the other hand, Brazier *et al*., reported a utility function with dead dummy following RUM[277], where no time

preference assumption was made. The function could be used with DCE data[4] and utility values were expected to be theoretically consistent with the DCE_{TTO} values[278]. This study used this framework for designing the DCE-Death question.

To test the varying effects engendered by distinct anchoring methodologies, a structured experimental analysis was conducted, encompassing two cohorts. Each cohort was tasked with engaging in different design tasks to enable a distinct type of anchoring: with duration or death.

Do the HWB preference differ for public samples from the UK and Australia?

On the EQ-HWB developing stage, it was found that general public and other stakeholders with varied culture background held different definition and preference for health and wellbeing[161]. This led to a concern if the divergence on HWB definition led to health state preference difference. EQ-5D observed systematic difference between value sets in different countries, and concluded culture background was the main influential factor on health state distribution and the scale of disutility[279]. Understanding and quantifying preference difference provided valuable insights for study design, i.e., prior value selection with efficient design, and HTA decisions on the value set choose.

While UK and Australia shared similar culture background, a similar preference on HWB was expected. This research would discern the variations in health preference among the general populations in the United Kingdom and Australia, and whether the preference difference influenced by design formats (DCE_{TTO} and DCE-Death).

Model specification and preference heterogeneity with HWB measure

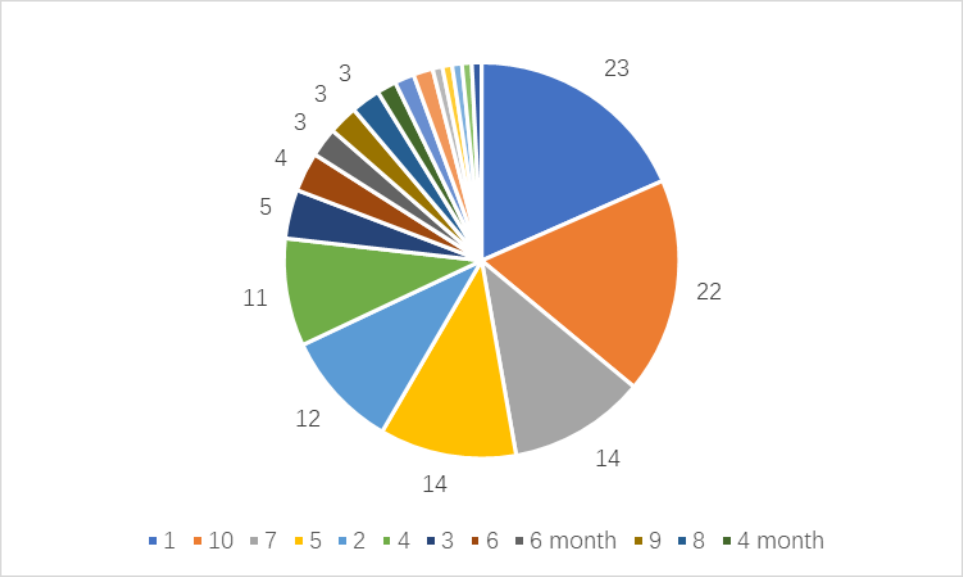
This study aimed to evaluate the performance of various utility models as well as data analysis approaches. The evaluated utility models were additive dummy coded model and the cross-attribute level effects model. It provided insight into which model best described the preference on HWB attributes and the model consistency itself.

6.2 Descriptive system

The health states were defined by 13 attributes of the EQ-HWB instrument (attribute selection see Chapter 4): Seeing, Hearing, Mobility, Daily activity, Control, Concentration/thinking clearly, Anxious, Sad/depressed, Loneliness, Support, Sleep, Fatigue, Pain Severity. For each attribute other than duration, five levels are used to describe the severity of impairment in monotonic order from 'No difficulty/None of the time' (level 1) to 'Unable/Most or all of the time' (level 5). For consistency and cognitive burden consideration, 7 of the 13 attributes in each choice set were overlapped to have the same level of severity in both of the compared health states presented[202]. The dimensions that were differed were highlighted with a light-yellow colour.

This study determined the duration levels by considering 1) literature review evidence, 2) qualitative evidence, and 3) optimal and orthogonal theories. According to the review evidence[123], duration levels 1, 2, 4, 5, 7, and 10 years were the most widely used (Figure 10). However, the qualitative consultation did not support including duration levels with short interval, and the "1, 2, 4" level combinations were rarely employed in other studies. The final selected duration levels were 1, 4, 5, 7, 10 years. The selected attributes were tested by focus group 3 and 4.

Figure 10 duration levels used in published literatures



6.3 Experimental design

As suggested by qualitative consultation and review works[129, 202], this feasibility study included three design formats: standard paired DCE_{TTO}, revised-order paired DCE_{TTO}, and triplet DCE-Death design (Table 1, Appendix C) to solve the research questions. The HWB attributes information in choice sets for all designs were the same for all three design formats. Standard and revised-order DCE_{TTO} designs incorporated duration attribute with levels 1, 4, 5, 7, 10 years. The triplet DCE-Death design included same 10-year duration attribute, and a triplet health state “immediate death”.

6.3.1 DCE Valuation task

The stan*dard paired DCE$_{TTO}$* was a paired comparison between two health states, where respondents were asked to choose between living in health state A and B in a fixed number of years. All of the HWB attributes were ordered in an EQ-HWB standard order (Table 1a, Appendix C).

The *revised-order paired* DCE_{TTO} format also presented paired choice, but the order of attributes was revised. The wellbeing attributes were presented first, followed by mental health and physical health attributes. A new attribute order was Control, Loneliness, Support, Concentration/thinking clearly, Anxious, Sad/depressed, Sleep, Fatigue, Seeing, Hearing, Mobility, Daily activity, Pain Severity and the duration

attribute (Table 1b, Appendix C). Information in each choice pair remained the same as *standard paired DCE$_{TTO}$* design.

The *triplet DCE-Death* format had three alternatives in each choice task, namely Health state A, B and C. Choice sets were presented as a series of triplet questions, consisting of 10 years in hypothetical health state A, B, and Health state C as dead state. The triplet design used a fixed duration information to prevent any duration preference interaction with attribute preference. Respondents were asked to choose which of the three options was the best and which was the worst. Each choice task thus provided the full ranking information of the three health states (Table 1c, Appendix C). The DCE-Death selections requested no time trade-offs [7, 8]. To compare with DCE$_{TTO}$, the *triplet DCE-Death* format did not change health and wellbeing attribute information for health state A and B.

This research did not consider any exclusion or prevention of 'implausible' level combinations, as the definition of implausible states were not consistent and an arbitrary exclusion led to inefficiency/non-orthogonality in the choice set selection[280].

6.3.2 Choice set selection

The 13 dimensions combined along with the duration attribute yielded over 1 billion health profiles, meaning it was necessary to include only a subset of profiles to include in the DCE. The DCE choice sets could be systematically selected using efficient design or generator design, as introduced in Chapter 1 and 3. Both were tested in this study.

Efficient design was constructed as pairs in Ngene[281]. A 'good' design provided the most significant amount of information with fixed number of choice sets[282]. The efficient design, with the criteria of lowest D-error[283], built using two prior assumptions: non-informative prior with nature order (i.e. this study assume the attribute levels had a nature order of coefficients, from 0.001 to 0.004, as we did not have estimates for all of the coefficients from previous studies), and average value for the 4 added attributes+9 UK pilot valuation coefficients. The experiment was designed

separately for paired and triplet comparison. The paired comparison experiments, with the duration attribute, was designed to allow estimates of main effects and two-factor interactions between each dummy-coded attribute level and the linear duration term, allowing collective estimation of main effects utility algorithm[44]. The triplet comparison experiments, with duration attribute level of 10-years, was designed with only main effects.

1,000,000 random pairs with 7 attributes overlapped were generated, as the candidate sets for Ngene to select the most efficient design using the modified Fedorov algorithm. The severity levels for each attribute were treated as categorical variable and duration levels as continuous variable in Ngene. A WTP approach was used for estimating the partial derivative disutility of paired comparison, derived using the modified Fedorov algorithm[281, 284, 285]. The number of choice tasks was 240 decided by the number of parameters to be estimated (14×4=56)[286], the sample size this survey expected to have, and the review suggestions[123, 287]. With the two prior assumptions reported above and two choice set formats (paired and triplet), there were 4 designs in total.

The generator design generated DCE design with the orthogonal array and generators[85, 99]. During the design stage, all attributes and levels were treated as dummy coded variables from 1 to the number of levels of any attribute. Then the orthogonal array was derived from the orthogonal list, and the generator was employed to produce the second option for each choice set. Unlike efficient design methods, the generator design did not necessitate precise prior values for error calculation. The generator design pairs with the HWB attributes were generated with R and Mathematica by Deborah Street. This design decided the number of pairs by the orthogonal array matrix[1], blocking rule and the number of overlapped attributes. By relaxing the balance requirement of blocking and allowing attributes appeared together as varied attribute in at least one generator, a 13 attribute HWB measure with 7 varied

[1] see the SAS webpage http://support.sas.com/techsup/technote/ts723_Designs.txt

attributes needed to have 6 generators.

Different from the efficient design, the triplet and paired formats had the same design with the HWB attributes in the choice sets. Generators for the HWB attributes were selected following criteria of information saturation level (how close it is to full factorial design)[85]. The duration information for paired formats was treated as a new attribute with 5 levels, added to the original generator by testing and comparing the efficiency performance of two-factor interactions[181,4,0], using a looping code in Mathematica. The generator design differed from the efficient design, as the efficient design allowed predetermined duration levels[85]. The interaction calculation and orthogonal arrays were influenced by the number of duration levels instead of the level itself. 600 choice sets were selected, with no dominated health states for the paired design and 48 were dominated health states for the triplet design. With the paired and triplet formats, there were two designs in total.

6.3.3 Simulation and evaluation

Simulation was conducted to test the design performance before deciding the choice set selection method. Simulation data were generated with two methods: either random answers (by Microsoft® Excel 'rand()' command), or use the prior information and normally distributed error term[167]. To keep the design and simulation process consistent, the efficient design prior information and information for generating simulation data remained the same. For the generator design, the information for simulation data generation was UK EQ-HWB-S valuation data for the 9 attributes + an average value for all other HWB attribute levels and duration. Following the random utility theory, the option with smaller disutility value was considered the selected option. The simulation data were modelled with conditional MNL model introduced by Burton *et al*.[288]. In total, eight simulation datasets (4 for efficient design and 4 for generator design) were created and modelled using the Stata 16 clogit command (Table 13).

The designs were evaluated by its performance with different prior information. First, with random data simulation, which was not considering any prior information, the

outcome of the conditional logit regression was expected to be insignificant for both the attribute level and the model.

Second, for the data simulation with prior information, where all of the respondents provided ‘correct’ answers, the models should be significant, having reasonable standard errors (S.D.), proper level coefficients, and predicting logically consistent values. Considering the available UK EQ-HWB-S valuation data, this study defined a level coefficient absolute value larger than 0.5 as high, and 0.8 as very high. Additionally, an S.D. larger than 5 and 10 times the coefficient was regarded as high and very high. Models predicting logically inconsistent values, where the higher levels had a lower disutility value than the lower levels or positive disutility values, would be problematic. The Akaike’s information criterion (AIC) and Schwarz’s Bayesian information criteria (BIC) were not considered, as the simulated observations and design used same amount of information[289, 290].

The simulation analysis reported log likelihood and pseudo R-squared number to evaluate the model performance. Pseudo R-squared is the equivalent statistic to R-squared statistic generated in ordinary least squares (OLS) regression that was often used for model goodness in logistic regressions. With the same design using a same amount of coefficient factors, the higher the log-likelihood value, the better a model explained the preference data[291]. The likelihood ratio (LR) chi-square evaluated the model significance. This study employed a 5% significance level for model evaluation, where the probability smaller than LR and pseudo R-squared value smaller or equal than 0.05 indicated the model was insignificant, or explained less variability than expected[292]. As introduced above, this research expected that the random model should be insignificant and explain none of the variability, while the model with simulation choices generated with prior values significant and explain much of the variability.

Table 13 Generator and efficient design and simulation summary

	Simulation 1 (S1)	Simulation 2 (S2)	Simulation 3 (S3)	Simulation 4 (S4)	Simulation 5 (S5)	Simulation 6 (S6)	Simulation 7 (S7)	Simulation 8 (S8)
	Generator design				Efficient design			
	Paired		Triplet		Paired		Triplet	
					Non-informative prior	with EQ-HWB prior	Non-informative prior	with EQ-HWB prior
Design summary	DCE_{TTO} design using the given generator		DCE design with the 13 selected attributes using the given generator		DCE_{TTO} design with standard order	DCE_{TTO} design with EQ-HWB-S prior	DCE design with standard order	DCE design with EQ-HWB-S prior
Pair number	600		600		240		240	
Simulate answers	600×150		600×150		240×250		240×250	
Selection	Orthogonality and statistical saturation				D-efficiency and orthogonality			
Simulation prior assumption	Non-informative prior	EQ-HWB-S UK valuation prior + random value	Non-informative prior	EQ-HWB-S UK valuation prior + random value	Non-informative prior	EQ-HWB-S UK valuation prior + random value	Non-informative prior	EQ-HWB-S UK valuation prior + random value

In comparison to efficient designs, generator designs emerged as a superior strategy for designing the EQ-HWB DCE survey (Table 14). First, in scenarios with non-informative priors with interaction model, both the generator design and the efficient design were insignificant with DCE_{TTO}. However, the efficient design with simulation 7 (S7) showed a significant model outcome with DCE-Death, with a chi-squred probability of 0.018 and R-squared value of 0.1213, suggesting that in the absence of proper simulation model, efficient designs might overestimate choice consistency. We also observed a larger number of disutility levels had lower disutility than the better levels, or had positive disutility values, with the efficient design. With the accurate prior information, the D-efficient design selected choice pairs with lower error variance, anticipating responses with low variance or uncertainty.

Nevertheless, deviations from this 'expected range' by respondents, or reliance on inaccurate prior information, could lead to model misestimation [89]. One study found that the complexity of choice questions in surveys led to less consistent responses from participants[90], which underscored the reliability of efficient design with informative priors from the EQ-HWB-S valuation outcome, or using a small sample pilot to update the efficient design for this valuation study.

Secondly, by assessing the efficacy of both designs incorporating EQ-HWB-S UK valuation priors, it was expected that the efficient design, enriched with prior information, would outperform the generator design in terms of model fitness. It was also expect the outperformance should be more significant with DCE-death simulation, as the EQ-HWB-S valuation adopted the EQ-VT valuation method[167]. However, observations deviated from these expectations. With the same model specification between efficient design and generator design, the Pseudo R-squared for two efficient designs were on the edge of insignificance, and the likelihood ratios were smaller than the generator design.

Yet, it was hard to explain why the R-squared of paired efficient design with duration interaction performed better than that of the triplet efficient design, where the prior values were calculated by main effect model instead of the interaction model used in DCE_{TTO}. A plausible hypothesis could be the impact of the MNL regression model, where incorporating duration led to a proportional reduction in coefficients, inadvertently boosting the model fitness value. Another explanation is that the interaction utility function and main effect utility function had little effect on the power of prior values.

In summary, the simulation results demonstrated that the generator design outperformed the efficient design with the EQ-HWB attributes, both when prior values were included and when they were not. Despite the generator design's lack of reliance on prior information, its inherent orthogonality and equal treatment of all elements seemed to enhance its capacity to predict main effects or interaction effects with duration[92]. Given the uncertainties associated with prior values, particularly due to the discrepancies between EQ-HWB-S and EQ-HWB attributes included, the generator design emerged as the preferred choice.

This study uses a generator design for both paired (DCE_{TTO}) and triplet (DCE-death) designs to select 600 choice sets per task design, with the method introduced in Section 6.3.2.

Table 14 generator and efficient design and simulation summary

	S1	S2	S3	S4	S5	S6	S7	S8
Design type	Generator design				Efficient design			
DCE format	DCE_{TTO}	DCE_{TTO}	DCE-death	DCE-death	DCE_{TTO}	DCE_{TTO}	DCE-death	DCE-death
Number of insignificant attribute level	49	1	51	3	52	7	43	12
Number of logical inconsistent attribute level	2	1	0	1	0	0	5	10
Number of high coefficients	NA	0	NA	5	NA	1	NA	2
Number of high S.D.	NA	0	NA	0	NA	3	NA	1
Likelihood ratio	-58809.2	-2087.28	-35303.1	-30295.5	-58821.3	-48612.2	-2730.47	-2055.06
Prob > chi2	0.117	<0.001	0.995	<0.001	0.878	<0.001	0.018	<0.001
Pseudo R-squared	<0.001	0.961	<0.001	0.133	<0.001	0.087	0.121	0.145
Performance	fit	fit	fit	fit	fit	fit	**Not fit**	fit

6.3.4 Blocking

As noted in Chapter 3, there were different options for selecting the choice sets that each participant answers, and a balance of attribute levels were always considered. This study implemented a BIBD within the generator design framework in selecting the choice sets, achieved by originating the generators from a BIBD structure in R. BIBD was considered on the design stage that if the BIBD had each pair of attributes appear together as non-zero entries in at least one generator, six generators were required (as the minimum number of generator). Ideally, the same number of choice sets were selected from each generator to form balanced blocks. Therefore, the decision was made to select 2 choice sets from each generator (2×6=12 choice sets for each block, 50 blocks in total), creating near-BIBD blocks for respondents. Several methods were considered for this selection process, including looping through code to allocate 2 choice sets from each generator until full distribution across 100 blocks, and starting the allocation process from specific number for 50 blocks. Systematic trials led to the identification of optimal starting points for each generator: 34, 28 (generator 1), 97, 1 (generator 2), 19, 55 (generator 3), 21, 46 (generator 4), 22, 10 (generator 5), and 76, 37 (generator 6).

The balanced blocking process resulted in 50 blocks, each with 12 choice sets, with or without dominated choice set. One extra choice set, from the 600 generator design pairs, was added to each of the 50 blocks to ensure each block had at least one dominated choice set, and the attribute levels were more balanced. As a result, each block had 13 tasks, where 12 choice sets were non-dominated, and one choice set was dominated (i.e. for every attribute had a level that was at least as good as the other profile). After the adjustment, each block had 13 choice sets and the level was balanced.

To evaluate the efficacy of each blocking design, statistical simulation tests and balance checks were conducted. Balance checks further scrutinized level distribution and representation within the choice sets. The final design comprised balanced blocks for both paired and triplet choices, consisting of 12 undominated and 1 dominated choice set presented in random order. Each block had no attribute level appeared more than 85 times and no less than 60 times (each block should have 13×2×13=338 piece of level information, equals to 338÷5=67.6 on average), allowing for comprehensive attribute level evaluation by each respondent. 82% of all blocks had a balanced level distribution. Blocking was completed separately for DCE_{TTO} and DCE-death designs. The design and the blocks were not updated after the pilot survey.

6.4 Survey design

The survey was administrated online. It began with an information sheet about the survey, followed by digital informed consent page. The survey had 4 parts: first, respondents completed demographic characteristics questions, health questions about disability/activity limitations and general health, and one repeated employment status question to check the data quality. The wording of questions for education level, income level, and ethnic groups were different for the two countries in accordance with the official classification. Respondents also completed the QoL (13 EQ-HWB attribute) questions, with all of the attributes were presented in the EQ-HWB measure standard order (See Appendix C).

Second, respondents were randomly allocated to one of the three DCE study arms until the target sample size was saturated and quota control satisfied for that arm. Respondents were shown instructions about the task and one practice DCE question that explained the question and gave feedback about their choice, enabling the respondents to confirm their paired/triplet comparison choice(s) or complete the practice question again. Respondents then completed the 13 DCE tasks in a random order (See Table 1, Appendix C).

Subsequently, respondents were asked a series of follow-up questions designed to delve their decision strategy, attribute importance questions, time preference questions and the death relative preference questions for triplet comparison samples. For the decision strategy question, respondents were asked about their consideration of specific attributes and how they made DCE choices. This aimed to uncover the cognitive strategies and value systems that participants employed when making their DCE choices. Next, the attribute importance question sought to identify the attributes that participants regarded as most crucial in their evaluations. The time preference question was adapted from the existing health economics literature on time preference, specifically designed to gauge their personal time preference on health[293]. For the DCE-Death respondents who never selected any health state as WTD, a specific death relative preference question was posed. This question sought to understand the reasons behind their reluctance to categorize any of the presented states as WTD.

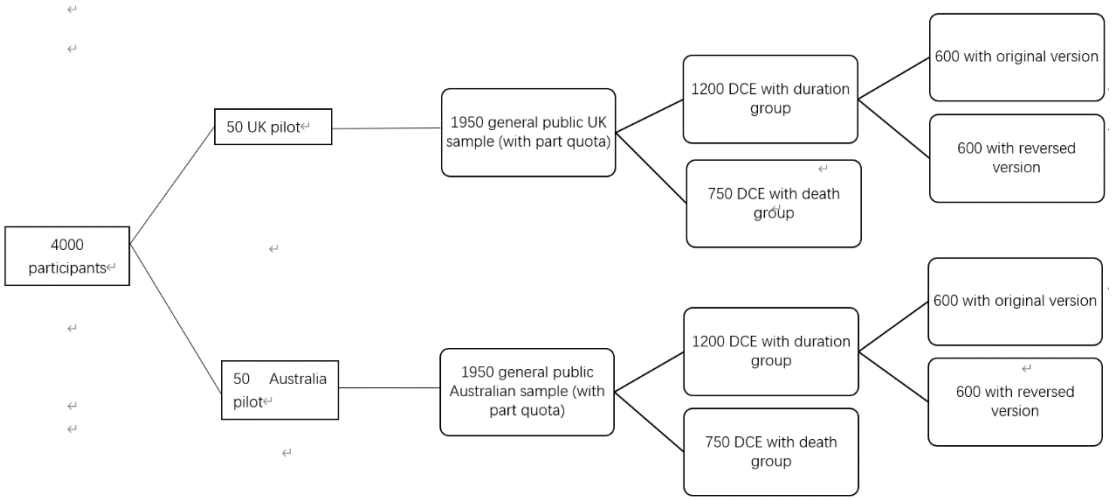
Finally, 5 multiple choice questions about how difficult the DCE tasks were to understand and answer, and whether the amount of information was proper for completing the tasks, were included along with a text question to understand respondents' opinion on the survey

and the content in general. Appendix C is the full survey questionnaire.

6.5 Selecting and recruiting the sample

Respondents were recruited using an existing online panel from a market research agency (the SurveyEngine GmbH Company) in both UK and Australia. The study aimed to recruit 2000 respondents in each country. The full samples were distributed into the three designs. The *standard paired* DCE_{TTO} format and *revised-order paired* DCE_{TTO} format had 600 participants, while the *triplet DCE-Death* format had 750 participants. A comprehensive summary of sample quotas for each design format can be seen in Figure 11.

Figure 11 sample in each country for each design format



Studies suggested that lower education levels can influence subjective feelings on DCE and understanding with EQ-HWB terms[294, 295]. Age was a significant factor in predicting their endurance for care[161, 296, 297], and gender was controlled in the majority of DCE studies[123]. The respondents were targeted to be representative in terms of age, gender, and education levels, with national census data from UK Office for National Statistics (ONS)[298], Australian Institute of Health and Welfare (AIHW)[299] and Australian Bureau of Statistics (ABS)[300-302] (Table 14). A tolerance of 5% variance was permitted for each group. A small incentive was provided if respondents completed the full requirements of the survey and passed the data quality check.

Table 14 Survey sample quota

	UK			Australia		
	Group	**Proportion**	**Adjusted**	**Group**	**Proportion**	**Adjusted**
Age	18-24	17%	16%	18-24	15%	15%
	25-34	17.10%	16.10%	25-34	17.43%	17.43%
	35-44	16.40%	15.40%	35-44	16.78%	16.78%
	45-54	16.80%	15.30%	45-54	15.60%	15.60%
	55-64	15.90%	14.40%	55-64	14.50%	14.50%
	65+	23.30%	22.30%	65+	21.10%	21.10%
Gender	Male	49%	49%	Male	49%	49%
	Female	51%	51%	Female	51%	51%
Education level	Primary school	27.80%	NA[1]	Year 11 and below	27.33%	7.30%
	Secondary school up to 16 years	13.40%	21.00%	Year 12	16.30%	20.30%
	Higher or secondary or further education (A-levels, BTEC, etc.)	22.20%	29.20%	Certificate	17.60%	21.60%
	College or university or professional qualification	22%	30%	Diploma	10%	14.20%
	Post-graduate degree	11.80%	19.80%	Bachelor	18.90%	22.90%
				Post-graduate	9.70%	13.70%

Note: the “Proportion” column indicates the national survey target population proportion. However, due to the characteristics of online DCE survey and the practical difficulty, people with Primary school/Year 11 and below education is hard to reach. An adjustment to the survey quota was implemented after discussing with the SurveyEngine. The “Adjusted” column reported the real survey quota for each group after adjustment.
1. The UK education policy request a compulsory education includes about 12 years until the age of 16. Respondents with primary education are hard to recruit and violate the policy. This research excluded these respondents. More can be found on.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/219167/v01-2012ukes.pdf

6.6 Piloting

The initial pilot launch recruited 50 individuals from each country (in total: 100) with the standard-order DCE_{TTO} and DCE-Death designs. Each design collected 25 respondents with no restrictions on the participant social-demographic characteristics. The UK pilot outcome was finished in December 2023, with all of the comparison data regressed using the conditional logit model to check the model significance. The proportion of respondents who reported difficulty on understanding the question and unconfident with their respondents were checked. The text feedback reporting any interpretation issues were summarized. The Australia pilot survey finished in March 2024 with the same analysis process conducted. The pilot did not reveal any major issues for revising the design and

choice set format.

6.7 Ethics approval

Ethics approval to conduct this research was granted by the University of Sheffield Population Health Ethics Panel (approval number: 051907) in the UK and Curtin University Human Research Ethics Office (approval number: HRE2024-0055) in Australia.

6.8 Data analysis

6.8.1 Data quality check

A data quality check was conducted by reviewing the repeated employment status question and the time used, to exclude those with inconsistent answer for the repeated question. Respondent IP was checked by the SurveyEngine and Haode Wang for excluding repeated answers or fraudulent data.

Further data quality assessments were performed for each cohort to check:

- Text feedback: all participant text comments were reviewed to categorize attitudes as negative (0), positive (1), or neutral (2), and the quantity of negative feedback was tallied.
- Survey time: responses with survey completion times under 10 minutes or any DCE question response times below 10 seconds were flagged.
- Left-right check: respondents who always chose left or right health state (cohort 1,2 and 3), or the middle health state as best/worst (cohort 3).
- Decision strategy and non-WTD check: the reason of respondents choosing death as the worst health state for all of the questions, and the proportion of these respondents were reported.

Note that these checks did not exclude any respondent (including the speeders), as it was hard to distinguish those with low data quality and true preference, but to report the general performance of respondents

6.8.2 Modelling with additive utility model

All of the data were analysed separately by design. DCE_{TTO} data was analysed using the conditional logit fixed effect model. The conditional logit model assumed that all of the

participants hold the same preference and each of the DCE choice was independent. For the standard-order paired DCE_{TTO} and revised-order paired DCE_{TTO} designs, an additive utility function with both time and EQ-HWB attribute levels was analysed consistently with the QALY approach[283]. The utility of the respondent *i* for the health state *j* in the choice task t is specified as[44]:

$$\mathrm{U}_{ij} = \beta_1 t_{ij} + \beta_2' X_{ij} \times t_{ij} + \varepsilon_{ij} \qquad \text{Formula (2)}$$

where the U_{ij} represents the utility of individual *i* for the health state *j*, of which the order and health state presented does not influence the utility of *j*. ε_{ij} represents the error term following the independent and identically distributed random variables (i.i.d) characteristic, β_1 and β_2' represent the coefficients for the duration in life years *t* and the coefficients on the 52 dummy-coded severity levels×duration, with level 1 (the best health attribute level) as the baseline. The β_1 represents the value respondent *i* assigns to living in perfect health for 1 year. The regression coefficients are latent values of each level interacted with duration, which have no utility interpretation.

Incorporating anchoring of latent coefficients with the structuring of choice tasks is a pivotal aspect of the study since the coefficients must be anchored on a QALY scale. The approach this study took here was attribute level coefficients differentiated with respect to coefficient of time. The anchored disutility β_2 of each level was:

$$\beta_2 = \frac{\beta_2^{'}}{\beta_1} \qquad \text{Formula (8)}$$

The DCE_{TTO} did not need explicit preference between health states and being dead, but a health state utility smaller than 0, achieved by using the function (9), was considered as WTD.

$$\frac{\mathrm{U}_{ij}}{t_{ij}} = \beta_1 + \beta_2' X_{ij} + \varepsilon_{ij} \qquad \text{Formula (9)}$$

The DCE-Death data were modelled using the conditional logit fixed effect model, mixed logit models and rank-ordered logit models. In theory, the rank-ordered model can be used to analyse the preference over a number of scenarios, with full ranking or partial ranking records[303]. A foundation for decomposing the triplet ranks to a series of paired comparison was proved by Craig, Buschbacher, and Salomon[304]. For the conditional logit model, the DCE-Death data was decomposed into three paired comparisons similar

to the DCE with duration. With the rank-order logit model, the full ranking of three health states can be pooled into the model for a complete ranking analysis. The hypothetical states were agreed with either the episodic or instant RUMs, as example within the UK MVH-protocol[278]. The utility function with both episodic and instant RUMs are:

$$\mathrm{U}_{ij(t)} = \begin{cases} \mu_j t + \varepsilon_{ij} & Episodic\ RUM \\ (\mu_j + \varepsilon_{ij})t = \mu_j t + \varepsilon_{ij} t & Instant\ RUM \end{cases} \quad \text{Formula (10)}$$

Where $\mathrm{U}_{ij(t)}$ is the total QALY in 10 years, μ_j is the state utility with a given HWB state, t is the 10-year time and ε_{ij} is error term. Taking the conclusion to logit model, under the linear time preference assumption with both utility and error term, function (11) had equal error term $\varepsilon_{ij} t = \varepsilon_{ij}$. We have the episodic RUM and instant RUM equal on a 10-year basis:

$$\mu_j t + \varepsilon_{ij} = \mu_j t + \varepsilon_{ij} t \quad \text{Formula (11)}$$

Thus, it is possible to calculate the HWB health state utility U_{ij} with instant RUM. The utility of the respondent *i* for the health state *j* in the choice task T is specified as[87]:

$$\mathrm{U}_{ij} = \beta_{1D} D_{ij} + \beta'_{2D} X_{ij} + \varepsilon_{ij} \quad \text{Formula (12)}$$

where U_{ij}, ε_{ij}, holds the same meaning but β_{1D} is the coefficient for the state of dead and β'_{2D} is the main effect of 52 dummy-coded severity levels. When D_{ij}=0, β_{1D} will not influence the units of interval attribute scales. The coefficients generated using Formula (12) is relative preference values not on a QALY scale. This research borrows a method from $\mathrm{DCE}_{\mathrm{TTO}}$[44], determining the relative value (distance) of non-dead health state attributes to death. The coefficients for each level β'_{2D} are divided by the coefficient for dead dummy β_{1D}, to get the disutility of each level on QALY scale β_{2D} (Formula 13). As β_{1D} and β'_{2D} are negative disutility values, the differentiation is the magnitude of coefficient contribution.

$$\widehat{\mathrm{U}_{ij}} = \frac{\beta_{1D} D_{ij}}{\beta_{1D}} + \frac{\beta'_{2D} X_{ij}}{|\beta_{1D}|} = 1 + \beta_{2D}$$

$$\beta_{2D} = \frac{\beta'_{2D}}{|\beta_{1D}|} \quad \text{Formula (13)}$$

The regression function assumed each parameter independently followed a normal distribution, as suggested by similar study[202]. The standard errors were calculated using

the Delta method with the Stata command *nlcom.* The regression outcomes reported in the following chapters, for the attribute levels, were anchored results. All of the results were not regressed with consistent model, where positive disutility, insignificant levels and non-monotonicity (where the higher levels had a lower disutility value than the lower levels) kept. Model performance was evaluated by the log-likelihood value, Pseudo R-squared statistics, AIC and BIC, as well as the inconsistent and insignificant coefficients[305]. All of the data analysis were completed in Stata version 17.

6.8.3 Models

A number of different models were estimated using data from each country:

Model 1 included all of the data from standard-order DCE_{TTO} cohort. The data was modelled with all of the level coefficients were dummy coded except the duration attribute. The independent terms were interaction terms consisting of duration and attribute level variables. This model is to generate a preference data for the first cohort. No preference heterogeneity considered. The utility function follows function (1).

Model 2 included all of the data from revised-order DCE_{TTO} cohort. Analysis was similar with Model 1.

Model 3 included all of the data from DCE-Death cohort. Each line of triplet comparison data was decomposed into three separate paired comparison replies, where the best health state was selected comparing with the rest two health states, and the worst health state was not selected comparing with the middle health state. The data was modelled with all of the level coefficients were dummy coded, including the death dummy representing the dead state. The utility function follows function (6).

Model 4 included all of the data from DCE-Death cohort. The data was modelled with the rank order logit model, with all of the level coefficients were dummy coded, including the death dummy representing the dead state. The utility function follows function (6).

6.8.4 Model performance and feasibility

The model reported the log likelihood statistic of the final logit regression model. The probability of likelihood ratio chi-square statistics (P-value) was used to evaluate the model significance. The pseudo R-square value, where measured the proportion of variance attribute to the predictors, was regarded as equivalent to the R-square in ordinary least square estimation[306]. A higher pseudo R-square value indicated better “goodness-to-fit”

for predicting the preference decisions.

The feasibility of each design strategy was compared and evaluated by the model significance first. A full feasibility was assessed using the proportion of left-right preferences, data quality, respondent understanding, response time, decision strategies, and feedback. This study intends to shed light on the importance of DCE task design and the feasibility comparison focuses on standard order DCE_{TTO} and DCE_{death}.

6.8.5 Order effect, design effect and country-level difference

The hypothesis that DCE preference algorithms tended to be influenced by order of information, design of choice set, and population were tested by comparing the values between analysis. To determine how the preference differ from each other, several key characteristics of the anchored coefficient were reported, including the number of logically inconsistent (non-monotonic and positive disutility) and insignificant attribute levels, the proportion of WTD health states, the location of the benchmark states (2222222222222, 3333333333333, 444444444444, worst state) on the value set scale[279], relative importance of health and wellbeing attributes, marginal effect of moving from one severity level to another, consistency of stated and regressed preference, coefficient difference significance, variance analysis and the health state value comparison.

A larger number of logically inconsistent and/or insignificant attribute level might be result of attribute wording, data quality and model performance. These numbers were compared between the three designs for each country. The attribute levels with significant but logically inconsistent anchored value demonstrated an understanding issue with the attribute level [124]. The regression outcome with smaller number of significant but logically inconsistent anchored coefficient was preferred.

The WTD proportion was assessed by conducting five million Monte Carlo simulations for each group. It was important to remember that the Monte Carlo simulated WTD proportion was different from listing all of the possible health states. However, the large number of possible EQ-HWB attribute combinations (5^{13}) prevented implementation of exhaustive method with any software package.

Differences in benchmark states between designs and countries were inspected by calculating the health state values for each design. A scale length was calculated by subtracting the worst state value from the perfect state value[279].

The relative importance of health and wellbeing attributes was evaluated by comparing the worst level disutility of physical attributes (Sleep, Fatigue, Seeing, Hearing, Mobility, Daily activity, Pain Severity) with wellbeing and mental health attributes (Control, Loneliness, Support, Concentration/thinking clearly, Anxious, Sad/depressed). The stated preference ranking and regressed attribute relative importance (by the attribute maximum value decrement) ranking were plotted and compared graphically.

The order effect comparison drew conclusion by comparing Model 1 and Model 2 attribute level significance, attributes preference order, relative preference between HWB attributes and the distribution of utility. This study conducted Wald test with all of the negative coefficients to test the significance of order effect (Model 1 and Model 2 results in each country) preference difference. The test was conducted with a null hypothesis that a particular pair of coefficients were equal, and calculated the weighed distance of two coefficients[307]. The P value was the probability that the null hypothesis was true, with a threshold of 0.05 and 0.1 to determine if the difference was significant. The difference was significant if the P value is smaller than the critical value.

A pooled conditional regression included all of the data from standard paired DCE_{TTO} and revised-order paired DCE_{TTO} cohorts was conducted. The data was modelled with all of the level coefficients were dummy coded except the duration attribute. The attribute order dummy $Order$ was coded as 0 if the choice set was in standard order, and 1 if it was in revised order. The utility function was as below.

$$\mathrm{U}_{ij} = \beta_1 t_{ij} + \beta_1 t_{ij} \times Order + \beta_2' X_{ij} \times t_{ij} + \beta_2' X_{ij} \times t_{ij} \times Order + \varepsilon_{ij} \qquad \text{Formula (14)}$$

A significant order effect term $\beta_1 t_{ij} \times Order$ revealed that respondents had varied preference with the attribute level.

The design effect compared Model 1 and Model 3 (standard order DCETTO regression and DCE-Death regression outcomes) anchored results. As the log likelihood ratio and Pseudo R-squared values are not comparable, the design effect compared number of non-monotonic or/and insignificant attribute levels, stated and regressed preference consistency, the distribution of utility values and the proportion of health states WTD.

Data quality was assessed by analyzing responses to dominated choice sets and consistency with the quality control question. Given the greater complexity of DCE_{death}

tasks, we anticipated longer response times and a higher proportion of reported understanding issues.

The modelled health state value distribution was analyzed to characterize value sets, including the proportion of WTD states and the overall range of values[25]. Due to the extensive number of possible HWB state combinations (5^{13}), we employed a Monte Carlo method with 108 simulations per design to estimate WTD proportions[26]. Utility values for a subset of states used in the valuation survey were calculated by mode and visualized. Health state values were compared using the ANOVA test, intraclass correlation coefficient (ICC) and Pearson's correlation to test if the paired values were significantly different and whether they changes with the same trend[17, 27]. Additionally, Wald tests were conducted to compare disutilities between levels 3 and 5, out of the consideration that the first disutility levels were expected to have positive disutility value and large proportion of non-monotonicity.

In sensitivity analyses, we excluded DCE_{death} respondents who did not classify any health state as WTD, and used this subset to generate value sets and calculate the WTD proportion, with comparisons to DCETTO values[308].

The comparison of preference in two countries included the examination of the ranking and relative importance of dimensions, relative decrements between levels, scale length difference and the distribution of utility, with Model 1 and Model 3 representing the DCE_{TTO} and DCE-Death results. Unlike doing Wald test with single anchored disutility, the country-level comparison focused on the relative preference of health and wellbeing, and the distribution of health state utilities. The specific utility values of 240 health states, selected by D-efficient method, and the 4 benchmark health states (111111111111, 2222222222222, 3333333333333 and 4444444444444) evaluated and compared two ways. First, the Model 1 (DCE_{TTO}) and Model 3 (DCE-Death) health state values in each country with the two value sets were plotted and compared. Second, a variance analysis with 240 utility values was conducted to evaluate the percentage of variance explained by sample source.

6.8.6 Modelling with CALE utility function

This research tested nonlinear cross-attribute level effect (CALE) model, introduced with EQ-5D-5L TTO data[309], hereafter referred to as "MULT16 (name the function as it had 13 level-5 attribute terms and 3 level effects – following the naming rule of cited journal

paper)"[310].

Unlike additive utility functions introduced in Section 7.3.1, the CALE model is a constrained main-effects model. It has 13 parameters are included representing the disutility of level 5 on each of the EQ-HWB attributes, and other three levels for each attribute are calculated by multiplying parameters for levels 4, 3 and 2 (L_4、L_3、L_2), with the assumption that level effect across the attributes are constant[311]. In other word, by using the funciton $\beta_5 \times L_2$, $\beta_5 \times L_3$, $\beta_5 \times L_4$, the disutility of level 2 – level 4 for each attribtue can be calculated. This constrained function relies on the assumption that the utility decrements for levels 2 to 4 are proportional to the level five utility decrement with the same parameter i.e. the proportional difference between the levels is fixed across all attributes but the coefficient of attribute level 5 (and hence levels 2 to 4 in proportion to this) varies. The DCE_{TTO} and DCE-Death had an extra term for duration main effect or dead dummy.

The assumption that participant's relative preferences on levels are independent is similar with the preference assumption of individual preference functions, but different from the assumption of additive DCE utility function[202], where the preference on each of the attribute level are independent of each other. The mathematical function of MULT16 is as follows for the DCE with duration data (more details can be found in Appendix D):

$$\mu_{ij} = \alpha + \beta_{1MT} t_{ij} + (\beta_{see5} See_{ij}^5 + \beta_{hr5} hr_{ij}^5 + \beta_{ar5} ar_{ij}^5 + \beta_{dtd5} dtd_{ij}^5 + \beta_{slp5} slp_{ij}^5 + \beta_{exh5} exh_{ij}^5 + \beta_{ll5} ll_{ij}^5 + \beta_{spt5} spt_{ij}^5 + \beta_{tk5} tk_{ij}^5 + \beta_{axs5} axs_{ij}^5 + \beta_{dpr5} dpr_{ij}^5 + \beta_{ctr5} ctr_{ij}^5 + \beta_{pp5} pp_{ij}^5) \times t_{ij} + (\beta_{see5} See_{ij}^4 + \beta_{hr5} hr_{ij}^4 + \beta_{ar5} ar_{ij}^4 + \beta_{dtd5} dtd_{ij}^4 + \beta_{slp5} slp_{ij}^4 + \beta_{exh5} exh_{ij}^4 + \beta_{ll5} ll_{ij}^4 + \beta_{spt5} spt_{ij}^4 + \beta_{tk5} tk_{ij}^4 + \beta_{axs5} axs_{ij}^4 + \beta_{dpr5} dpr_{ij}^4 + \beta_{ctr5} ctr_{ij}^4 + \beta_{pp5} pp_{ij}^4) L_4 \times t_{ij} + (\beta_{see5} See_{ij}^3 + \beta_{hr5} hr_{ij}^3 + \beta_{ar5} ar_{ij}^3 + \beta_{dtd5} dtd_{ij}^3 + \beta_{slp5} slp_{ij}^3 + \beta_{exh5} exh_{ij}^3 + \beta_{ll5} ll_{ij}^3 + \beta_{spt5} spt_{ij}^3 + \beta_{tk5} tk_{ij}^3 + \beta_{axs5} axs_{ij}^3 + \beta_{dpr5} dpr_{ij}^3 + \beta_{ctr5} ctr_{ij}^3 + \beta_{pp5} pp_{ij}^3) L_3 \times t_{ij} + (\beta_{see5} See_{ij}^2 + \beta_{hr5} hr_{ij}^2 + \beta_{ar5} ar_{ij}^2 + \beta_{dtd5} dtd_{ij}^2 + \beta_{slp5} slp_{ij}^2 + \beta_{exh5} exh_{ij}^2 + \beta_{ll5} ll_{ij}^2 + \beta_{spt5} spt_{ij}^2 + \beta_{tk5} tk_{ij}^2 + \beta_{axs5} axs_{ij}^2 + \beta_{dpr5} dpr_{ij}^2 + \beta_{ctr5} ctr_{ij}^2 + \beta_{pp5} pp_{ij}^2) L_2 \times t_{ij} + \epsilon_n$$

Formula (15)

where the U_{ij} represents the utility of individual *i* for the health state *j*, of which the order and health state presented did not influence the utility of *j*. ϵ_n represents the error term, β_{1MT} and β_{see5}- β_{pp5} represent the coefficients of 13 dummy-coded level 5×duration for

the attributes, compared to level 1 (the best health attribute level). The $See_{ij}^{5} - pp_{ij}^{5}$ are the dummy-coded level 5. L_4、L_3、L_2 represents the proportional factor for level 4 to level 2. To anchor the coefficients of level 5 onto a 0 to 1 utility scale, the coefficients are divided by the coefficient for dead dummy main effect β_1.

The model specification for the DCE-Death data is:

$$\mu_{ij} = \beta_{1MD}Dummy_{death} + (\beta_{see5}See_{ij}^{5} + \beta_{hr5}hr_{ij}^{5} + \beta_{ar5}ar_{ij}^{5} + \beta_{dtd5}dtd_{ij}^{5} + \beta_{slp5}slp_{ij}^{5} + \beta_{exh5}exh_{ij}^{5} + \beta_{ll5}ll_{ij}^{5} + \beta_{spt5}spt_{ij}^{5} + \beta_{tk5}tk_{ij}^{5} + \beta_{axs5}axs_{ij}^{5} + \beta_{dpr5}dpr_{ij}^{5} + \beta_{ctr5}ctr_{ij}^{5} + \beta_{pp5}pp_{ij}^{5}) + (\beta_{see5}See_{ij}^{4} + \beta_{hr5}hr_{ij}^{4} + \beta_{ar5}ar_{ij}^{4} + \beta_{dtd5}dtd_{ij}^{4} + \beta_{slp5}slp_{ij}^{4} + \beta_{exh5}exh_{ij}^{4} + \beta_{ll5}ll_{ij}^{4} + \beta_{spt5}spt_{ij}^{4} + \beta_{tk5}tk_{ij}^{4} + \beta_{axs5}axs_{ij}^{4} + \beta_{dpr5}dpr_{ij}^{4} + \beta_{ctr5}ctr_{ij}^{4} + \beta_{pp5}pp_{ij}^{4})L_4 + (\beta_{see5}See_{ij}^{3} + \beta_{hr5}hr_{ij}^{3} + \beta_{ar5}ar_{ij}^{3} + \beta_{dtd5}dtd_{ij}^{3} + \beta_{slp5}slp_{ij}^{3} + \beta_{exh5}exh_{ij}^{3} + \beta_{ll5}ll_{ij}^{3} + \beta_{spt5}spt_{ij}^{3} + \beta_{tk5}tk_{ij}^{3} + \beta_{axs5}axs_{ij}^{3} + \beta_{dpr5}dpr_{ij}^{3} + \beta_{ctr5}ctr_{ij}^{3} + \beta_{pp5}pp_{ij}^{3})L_3 + (\beta_{see5}See_{ij}^{2} + \beta_{hr5}hr_{ij}^{2} + \beta_{ar5}ar_{ij}^{2} + \beta_{dtd5}dtd_{ij}^{2} + \beta_{slp5}slp_{ij}^{2} + \beta_{exh5}exh_{ij}^{2} + \beta_{ll5}ll_{ij}^{2} + \beta_{spt5}spt_{ij}^{2} + \beta_{tk5}tk_{ij}^{2} + \beta_{axs5}axs_{ij}^{2} + \beta_{dpr5}dpr_{ij}^{2} + \beta_{ctr5}ctr_{ij}^{2} + \beta_{pp5}pp_{ij}^{2})L_2 + \epsilon_n \quad \text{Formula (16)}$$

where U_{ij}, ε_{ij}, L_4、L_3、L_2 holds the same meaning but β_{1MD} is the coefficient for the state of dead and β_{see5}- β_{pp5} represent the coefficients of 13 dummy-coded level 5 main effect. To anchor the coefficients onto a 0 to 1 utility scale, the coefficients for each level β_{see5}- β_{pp5} are divided by the coefficient for dead dummy β_{1D}, which is the disutility of level 5 anchored onto the scale.

6.8.7 Robustness

Robustness of the results was examined by estimating the same models excluding respondents who provided lower quality responses. This included: respondents who did not choose the dominant profile in the dominance questions; respondents who reported it was difficult to understand the survey (very or slight understanding issue); and respondents who spent time shorter than one quarter of average time on answering the DCE questions, including answering the practice question.

6.8.8 Preference heterogeneity

Analysing preference heterogeneity forms the core part of this feasibility study, which aimed to discern the impact of observable and unobservable preference heterogeneity. There was a growing number of studies accounting the preference heterogeneity in the

DCE data analysis[312], to provide richer implementation of the valuation preference data. The preference heterogeneity analysis considered the ISPOR recommended preference heterogeneity analysis framework[313], including explain preference heterogeneity with the interaction terms (i.e., interaction terms with the key personal characteristics), accounting for unexplained heterogeneity with continuous distribution (mixed logit estimation), and with latent class estimations. Due to the small sample size with each cohort and large number of attributes, the mixed logit estimation was infeasible. All of the preference heterogeneity analysis was conducted with the respondents of standard-order DCE_{TTO} design

Regression with interaction terms was to include interactions between an attribute variable and the observed characteristic of the respondent. A significant interaction term indicated the preference heterogeneity existed between the known groups (Function 17). The $\gamma_{interaction}$ is the individual characteristic in each regression and the other factors are same as the Formula 2.

$$\mathrm{U}_{ij} = \beta_1 t_{ij} + \beta_1 t_{ij} \times \gamma_{interaction} + \beta_2' X_{ij} \times t_{ij} + \beta_2' X_{ij} \times t_{ij} \times \gamma_{interaction} + \varepsilon_{ij} \quad \text{Formula (17)}$$

A second analysis was latent class analysis, which classified respondents with similar decision patterns to various latent classes[267]. The utility function allowing preference heterogeneity was:

$$\mathrm{U}_{ij} = \beta_1 t_{ij} + F_n t_{ij} \times \delta_{latent} + \beta_2' X_{ij} \times t_{ij} + F_m' X_{ij} \times t_{ij} \times \delta_{latent} + \varepsilon_{ij} \quad \text{Formula (18)}$$

where the F_n is the matrix of effect of latent class factors and δ_{latent} is a column vector of the latent class factors. The other terms are the same as the Formula 2[314]. Latent class regression produces parameter estimations of each class and the class share of each class from the overall dataset. The optimum number of classes was determined using the Corrected AIC (CAIC) and BIC criteria, testing 2 to 10 classes for UK and Australian dataset separately. The class model with the lowest CAIC and BIC was preferred.

However, the ISPOR studies did not discuss individual characteristics that should be included, especially examined little evidence of WTP studies. Individual characteristics included in this study were age (18-65 years old and 65 or above), gender (female or male), carer (informal carer of adult or not), cared (cared by other people or not) and health conditions (good health or poor health).

6.9 Conclusion

This chapter outlined the methodological framework for study design and data analysis. By utilizing simulation data, the study selected the generator design approach to design the proposed DCE surveys. Compared with similar studies, the design method selection process was more systematic and comprehensive. Further refinement in the blocking of choice sets was achieved through the application of balanced approach, in the perspective of attribute level and information provided in each block. This systematic design was not only conducive to a more engaging respondent experience but also significantly improve the statistical efficiency. The data analysis was aligned with the study's research objectives, with a particular emphasis on evaluating the influential factors in a comprehensive health and wellbeing measure. This study design was both scientifically robust and highly relevant to the study's objectives.

Chapter 7 DCE analysis results

This chapter reports the analysis result. The DCE survey was generally well implemented, and the data analysis were able to derive a consistent and meaningful result with the conditional logit model.

Following the study design reported in Chapter 6, the survey was implemented with three designs: DCE_{TTO} with standard order of information (Design 1), DCE_{TTO} with revised-order of information (Design 2) and DCE-Death (Design 3). This Chapter reported general sample characteristics, comparison results of order of information, survey design and preference heterogeneity in two countries, and the CALE model regression outcome. The preference heterogeneity was also discussed.

7.1 The sample

7.1.1 The UK sample

The demographic and health characteristics of the sample were presented in Table 15. The survey was completed by 2037 respondents, where 34 of them were excluded due to bot check or repeated IP. 2003 respondents were included in the final analysis after the data quality check. The sample was generally representative of UK population in age, gender and marital status. Approximately 60% of all respondents had a university equivalent degree or higher, more highly educated than the UK population. The sample had 56%, 21 % and 7% of respondents employed (full-time or part-time), retired or students, which were the three main employment status. Around 51% of respondents earned an income higher than the median level of UK household income[2]. 24% were parents or guardians for child or children aged under 18. 14% were informal carers taking care of adult family member or friend as informal carer, while 7% were being cared by an informal carer.

Samples demographic in each group was well balanced, except standard order DCE_{TTO}

[2] This number may be different from varied sources. This research took the value from 2023 UK Family Resources Survey (FRS), full report available at https://www.gov.uk/government/statistics/households-below-average-income-for-financial-years-ending-1995-to-2023/households-below-average-income-an-analysis-of-the-uk-income-distribution-fye-1995-to-fye-2023. The median household income is £32,500 per year.

respondents were slightly older and had higher retirement proportion (1.5% higher than average). DCE-Death respondents had larger proportion (2.3% higher than average) of respondents with bachelor's or above education (Table 15).

Table 15 the descriptive characteristics of UK data

Characteristics	General population [1]	Design 1[2]		Design 2[2]		Design 3[2]		Overall	
		No.	%	No.	%	No.	%	No.	%
		627		600		776		2003	
Sex									
Male	49%	306	48.80%	288	48.00%	381	49.10%	975	49.20%
Female	51%	321	51.20%	311	51.83%	394	50.77%	1023	50.27%
Prefer not to say		0	0	1	0.17%	1	0.12%	2	0.09%
Age									
Ave. age	40.70	47.47		46.92		46.03		46.74	
Education									
Primary school	27.80%	3	0.48%	2	0.33%	1	0.13%	6	0.30%
Secondary school up to 16 years	13.40%	118	18.82%	112	18.67%	137	17.65%	367	18.32%
Higher or secondary or further education (A-levels, BTEC, etc.)	22.20%	142	22.65%	149	24.83%	163	21.01%	454	22.67%
College or university or professional qualification	22%	244	38.92%	235	39.17%	326	42.01%	805	40.19%
Post-graduate degree	11.80%	118	18.82%	99	16.50%	146	18.81%	363	18.12%
Prefer not to say		2	0.31%	1	0.16%	5	0.64%	8	0.40%
Marital status									
Single	34.5%	238	37.96%	208	34.67%	266	34.28%	712	35.55%
Married/ Living with partner	50.3%	313	49.92%	340	56.67%	439	56.57%	1,092	54.52%
Separated/ Divorced	9.1%	51	8.13%	34	5.67%	48	6.19%	133	6.64%
Widowed	6.1%	22	3.51%	16	2.67%	21	2.71%	59	2.95%
Prefer not to say		3	0.48%	2	0.33%	2	0.26%	7	0.35%
Average time									
Total (seconds)		990.84		983.88		1122.45		1039.74	
S.D.		789.84		729.48		1275.48		994.18	
DCE question (seconds)		28.46		27.24		31.24		29.17	

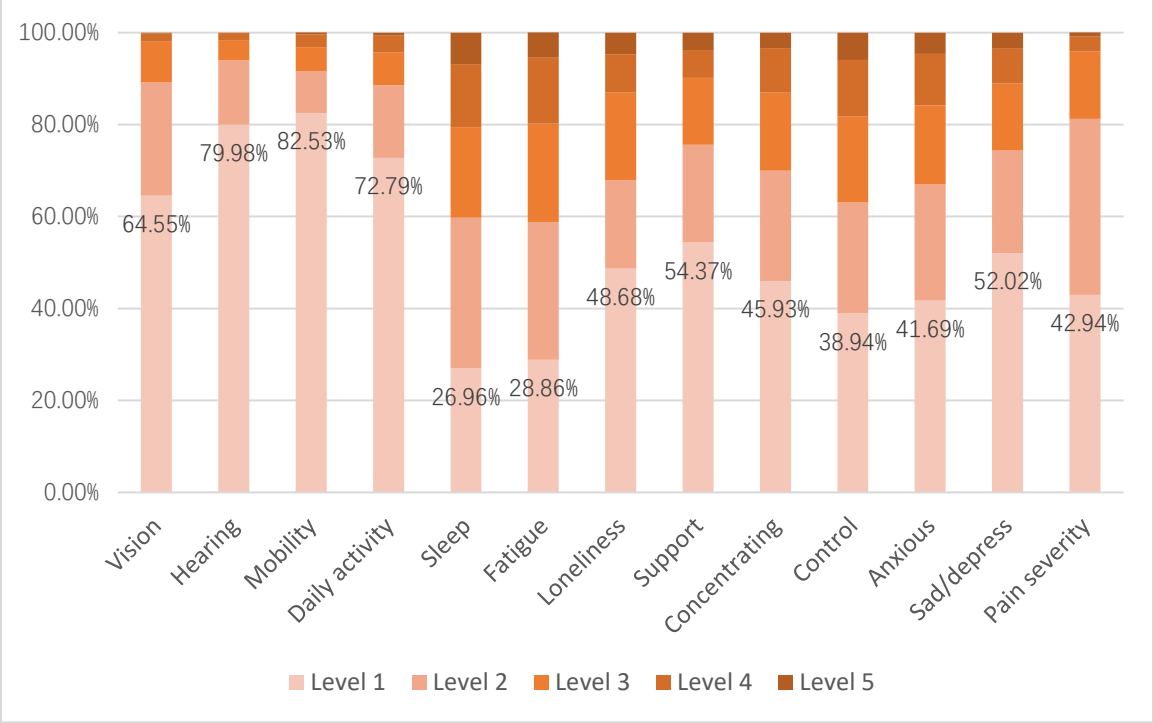
S.D.	39.56		28.08		32.14		33.58	
Employment								
Full-time employed or self-employed	242	38.60%	245	40.83%	342	44.07%	829	41.39%
Part-time employed or self-employed	104	16.59%	101	16.83%	115	14.82%	320	15.98%
Retired	143	22.81%	124	20.67%	164	21.13%	431	21.52%
Student	39	6.22%	50	8.33%	52	6.70%	141	7.04%
Unemployed	52	8.29%	29	4.83%	44	5.67%	125	6.24%
Long-term sickness	21	3.35%	15	2.50%	25	3.22%	61	3.05%
Look after family/home	20	3.19%	30	5.00%	27	3.48%	77	3.84%
Prefer not to say	1	0.16%	1	0.17%	1	0.13%	3	0.15%
Other	5	0.80%	5	0.83%	6	0.77%	16	0.80%
Income level								
Up to £5,199	27	4.31%	18	3.00%	28	3.61%	73	3.64%
£5,200 and up to £10,399	37	5.90%	24	4.00%	22	2.84%	83	4.14%
£10,400 and up to £15,599	48	7.66%	35	5.83%	52	6.70%	135	6.74%
£15,600 and up to £20,799	42	6.70%	50	8.33%	53	6.83%	145	7.24%
£20,800 and up to £25,999	61	9.73%	64	10.67%	79	10.18%	204	10.18%
£26,000 and up to £31,199	78	12.44%	75	12.50%	91	11.73%	244	12.18%
£31,200 and up to £36,399	47	7.50%	45	7.50%	63	8.12%	155	7.74%
£36,400 and up to £51,999	106	16.91%	120	20.00%	175	22.55%	401	20.02%
£52,000 and above	136	21.69%	141	23.50%	171	22.04%	448	22.37%
Prefer not to say or don’t know	45	7.18%	28	4.67%	42	5.41%	115	5.74%
Parent guardian								
Parent or guardian for a child or children aged under 18 years	137	21.85%	149	24.83%	200	25.77%	486	24.26%
Not a parent or guardian for a child or children aged under 18 years	489	77.99%	450	75.00%	575	74.10%	1,514	75.59%
Prefer not to say	1	0.16%	1	0.17%	1	0.13%	3	0.15%
Care status								

Carer for an adult(s) family member or friend (not as a paid job)	94	14.99%	88	14.67%	105	13.53%	287	14.33%
Cared for by other adults (including paid carers) because of health or age	45	7.18%	43	7.17%	52	6.70%	140	6.99%
Neither of the above	479	76.40%	464	77.33%	612	78.87%	1555	77.63%
Prefer not to say or don't know	9	1.44%	5	0.83%	7	0.90%	21	1.05%
Day-to-day activities limitation								
Yes, limited a lot	50	7.97%	40	6.67%	50	6.44%	140	6.99%
Yes, limited a little	117	18.66%	127	21.17%	156	20.10%	400	19.97%
No	455	72.57%	428	71.33%	565	72.81%	1,448	72.29%
Prefer not to say	5	0.80%	5	0.83%	5	0.64%	15	0.75%
General health								
Excellent	59	9.41%	71	11.83%	94	12.11%	224	11.18%
Very good	208	33.17%	186	31.00%	229	29.51%	623	31.10%
Good	196	31.26%	194	32.33%	278	35.82%	668	33.35%
Fair	141	22.49%	124	20.67%	147	18.94%	412	20.57%
Poor	23	3.67%	25	4.17%	28	3.61%	76	3.79%

Notes: *9 people did not provide the sex, (and/or) education, (and/or) marital status information.
[1]The general population quota see Chapter 6.5 for more detailed information.
[2]Desgin 1: DCE_{TTO} design with the EQ-HWB attribute order (health first); Design 2: DCE_{TTO} design with the revised attribute order (wellbeing first); Design 3: DCE-Death design.

Overall, 27% respondents self-reported limitation with day-to-day activities, and 24% classified them as fair or poor health. The proportion was 1% and 3% higher for the DCE_{TTO} groups (Table 15). The EQ-HWB (13 attributes) responses demonstrated the health problems encountered by the sample in the last 7 days, where over half of all respondents reported sleep, exhausted and tiredness, anxious, depression, and physical pain (Figure 12). The respondents were distributed primarily across the first two severity levels of the physical health attributes, and the first three of the wellbeing and mental health attributes. No significant variance identified among the groups by conducting one-way ANOVA test, except with Concentration/thinking clearly and Depression attributes that had a variance of 5%-6% with the first two levels (Table 1, Appendix E).

Figure 12 EQ-HWB self-report of UK respondents



Note: Level 1 is No difficulty or None of the time; Level 2 is Slight difficulty or Only occasionally; Level 3 is Some difficulty or Sometimes; Level 4 is A lot of difficulty or Often; Level 5 is Unable or Most or all of the time. Pain severity levels are No physical pain (Level 1), Mild physical pain (Level 2), Moderate physical pain (Level 3), Severe physical pain (Level 4), Very severe (Level 5). Percentage of Level 1 is provided as an example for compare the ceiling effect. More details for the other percentages reported in the Table 1, Appendix E.

7.1.2 The Australian sample

The study recruited 2068 participants through four rounds of invitations. 49 respondents were excluded. Characteristics of the 2019 participants included after data quality check reports in Table 16. Although we sought to recruit a representative of the Australian

population, the included samples were older and well-educated: average age of the sample was 46.7 (versus average 40.7) and 44% had higher or above education (versus 29% in the Australia general population). 46%,17% and 19% of all respondents had full-time job, part-time job or were retired. Roughly a quarter of all the respondents were carers and 14% were cared for by other people. 39% of respondents earned an income higher than the median level[3]. 31% were guardians for child or children aged under 18. 24% were informal carers taking care of adult family members or friends as informal carer, while 14% were being cared by an informal carer, higher proportion than the UK sample.

The DCE-Death group recruited more female, younger, single and well-educated respondents than the two DCE_{TTO} groups. The employment, income and care status were well-balanced (Table 16).

[3] Here we cited the government household wealth data
https://www.abs.gov.au/statistics/economy/finance/household-income-and-wealth-australia/latest-release
The median income is $121,108 for the latest financial year 2019-2020

Table 16 The descriptive characteristics of full Australian data

Characteristics	General population [1]	Design 1[2]		Design 2[2]		Design 3[2]		Overall	
		No.	%	No.	%	No.	%	No.	%
		632		604		783		2019	100.00%
Sex									
Male	49%	325	51.42%	308	50.99%	353	45.08%	986	48.84%
Female	51%	306	48.42%	294	48.68%	427	54.53%	1027	50.87%
Prefer not to say		1	0.16%	2	0.33%	3	0.38%	6	0.30%
Age									
Ave. age	40.70	48.56		47.96		44.13		46.74	
Education									
Year 11 and below	27.33%	81	12.82%	64	10.60%	78	9.96%	223	11.05%
Year 12	16.30%	87	13.77%	103	17.05%	122	15.58%	312	15.45%
Certificate (any level including trade certificate)	17.60%	123	19.46%	98	16.23%	119	15.20%	340	16.84%
Diploma/ advanced diploma	10%	71	11.23%	72	11.92%	103	13.15%	246	12.18%
Bachelors or honours degree	18.90%	181	28.64%	184	30.46%	264	33.72%	629	31.15%
Post-graduate degree (Masters or Doctorate)	9.70%	86	13.61%	83	13.74%	94	12.01%	263	13.03%
Prefer not to say		3	0.47%	0	0.00%	3	0.38%	6	0.30%
Marital status									
Single	33%	163	25.79%	171	28.31%	252	32.18%	586	29.02%
Married/ Living with partner	58%	384	60.76%	353	58.44%	439	56.07%	1176	58.25%
Separated/ Divorced	9%	66	10.44%	47	7.78%	63	8.05%	176	8.72%
Widowed	6%	15	2.37%	31	5.13%	22	2.81%	68	3.37%
Prefer not to say		4	0.63%	2	0.33%	7	0.89%	13	0.64%
Average time									
Total (seconds)		1148.64		1094.46		1212.17		1157.07	
S.D.		828.59		730.49		894.22		828.47	

DCE question (seconds)	27.51		28.44		29.23		28.46	
S.D.	27.68		27.79		30.11		28.67	
Employment								
Full-time employed or self-employed	278	43.99%	280	46.36%	363	46.36%	921	45.62%
Part-time employed or self-employed	102	16.14%	89	14.74%	154	19.67%	345	17.09%
Retired	140	22.15%	139	23.01%	110	14.05%	389	19.27%
Student	24	3.80%	19	3.15%	53	6.77%	96	4.75%
Unemployed	42	6.65%	41	6.79%	50	6.39%	133	6.59%
Long-term sickness	15	2.37%	12	1.99%	18	2.30%	45	2.23%
Look after family/home	18	2.85%	19	3.15%	24	3.07%	61	3.02%
Prefer not to say	3	0.47%	0	0.00%	1	0.13%	4	0.20%
Other	10	1.58%	5	0.83%	10	1.28%	25	1.24%
Income level								
Negative or Zero Income	11	1.74%	6	0.99%	6	0.77%	23	1.14%
$1 - $20,799 per year ($1 - $399 per week)	29	4.59%	30	4.97%	28	3.58%	87	4.31%
$20,800 - $41,599 per year ($400 - $799 per week)	103	16.30%	110	18.21%	123	15.71%	336	16.64%
$41,600 - $77,999 per year ($800 - $1499 per week)	147	23.26%	116	19.21%	172	21.97%	435	21.55%
$78,000 - $103,999 per year ($1500 - $1999 per week)	84	13.29%	77	12.75%	127	16.22%	288	14.26%
$104,000 - $155,999 per year ($2000- $2999 per week)	93	14.72%	106	17.55%	130	16.60%	329	16.30%
$156,000 - $207,999 per year ($3000 - $3999 per week)	44	6.96%	48	7.95%	73	9.32%	165	8.17%
$208,000 - $259,999 per year ($4000 - $4999 per week)	31	4.91%	27	4.47%	32	4.09%	90	4.46%
$260,000 - $311,999 per year ($5000 - $5999 per week)	43	6.80%	36	5.96%	36	4.60%	115	5.70%
$312,000 or more per year ($6000 or more per week)	20	3.16%	27	4.47%	29	3.70%	76	3.76%

Prefer not to say or don't know	27	4.27%	21	3.48%	27	3.45%	75	3.71%
Parent guardian								
Parent or guardian for a child or children aged under 18 years	176	27.85%	195	32.28%	249	31.80%	620	30.71%
Not a parent or guardian for a child or children aged under 18 years	452	71.52%	408	67.55%	530	67.69%	1390	68.85%
Prefer not to say	4	0.63%	1	0.17%	4	0.51%	9	0.45%
Care status								
Carer for an adult(s) family member or friend (not as a paid job)	151	23.89%	153	25.33%	185	23.63%	489	24.22%
Cared for by other adults (including paid carers) because of health or age	91	14.40%	87	14.40%	111	14.18%	289	14.31%
Neither of the above	12	1.90%	5	0.83%	13	1.66%	30	1.49%
Prefer not to say or don't know	378	59.81%	359	59.44%	474	60.54%	1211	59.98%
Day-to-day activities limitation								
Yes, limited a lot	59	9.34%	59	9.77%	77	9.83%	195	9.66%
Yes, limited a little	184	29.11%	149	24.67%	199	25.42%	532	26.35%
No	384	60.76%	394	65.23%	501	63.98%	1,279	63.35%
Prefer not to say	5	0.79%	2	0.33%	6	0.77%	13	0.64%
General health								
Excellent	81	12.82%	83	13.74%	127	16.22%	291	14.41%
Very good	165	26.11%	174	28.81%	223	28.48%	562	27.84%
Good	221	34.97%	204	33.77%	254	32.44%	679	33.63%
Fair	140	22.15%	112	18.54%	140	17.88%	392	19.42%
Poor	25	3.96%	31	5.13%	39	4.98%	95	4.71%

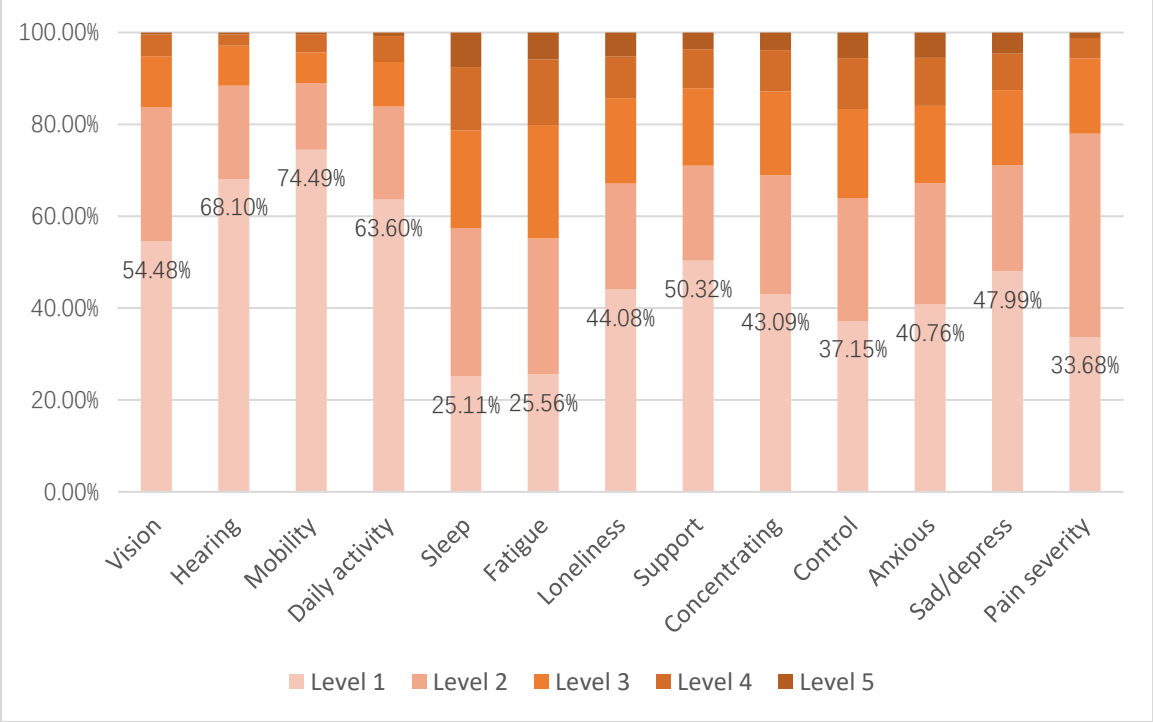
Notes: *13 people did not provide the gender, (and/or) education, (and/or) marital status information

[1]The general population quota see Chapter 6.5 for more detailed information.

[2]Desgin 1: DCE_{TTO} design with the EQ-HWB attribute order (health first); Design 2: DCE_{TTO} design with the revised attribute order (wellbeing first); Design 3: DCE-Death design.

42% of respondents had very good health and 34% had good health, while 36% of respondents reported little to serious limitation with their daily activities. The self-report health of Australian sample was worse than the UK sample, by a higher proportion of poor health. Among the three groups, the standard-order DCE_{TTO} respondents had higher proportion of Daily activity limitation but reported better general health. In general, the ANOVA test indicated that Daily activity limitation and general health had no significant difference (Table 2, Appendix E). The EQ-HWB self-reported measure recorded sleep (75%), exhausted and tiredness (75%), physical pain (67%), anxiety (63%) and depression (59%) as the most common health issues (Figure 12). The ANOVA test result indicated that except for the Day-to-day activities attribute, DCE_{TTO} respondents were better with almost all wellbeing attributes, including Lonely, support, Concentration/thinking clearly, Anxious, Depression and Control, with 4% to 12% higher proportion of level 1 and level 2 report (Table 2, Appendix E).

Figure 12: EQ-HWB self-report of Australian respondents



Note: Level 1 is No difficulty or None of the time; Level 2 is Slight difficulty or Only occasionally; Level 3 is Some difficulty or Sometimes; Level 4 is A lot of difficulty or Often; Level 5 is Unable or Most or all of the time. Pain severity levels are No physical pain (Level 1), Mild physical pain (Level 2), Moderate physical pain (Level 3), Severe physical pain (Level 4), Very severe (Level 5). Percentage of Level 1 is provided as an example for compare the ceiling effect. More details for the other percentages reported in the Table 1, Appendix E.

7.2 Understanding and data quality check

This section reported the time for completing the EQ-HWB DCE valuation study, confidence with the answers and data quality checks. Data quality assessments considered the text feedback, the repeated employment status question answer (asked twice at different points during the survey to assess consistency in responses), survey completion time, respondents who always selected the left or right answer, respondent decision pattern and the proportion of respondents who did not select any state as being WTD due to the varied anchoring method reason.

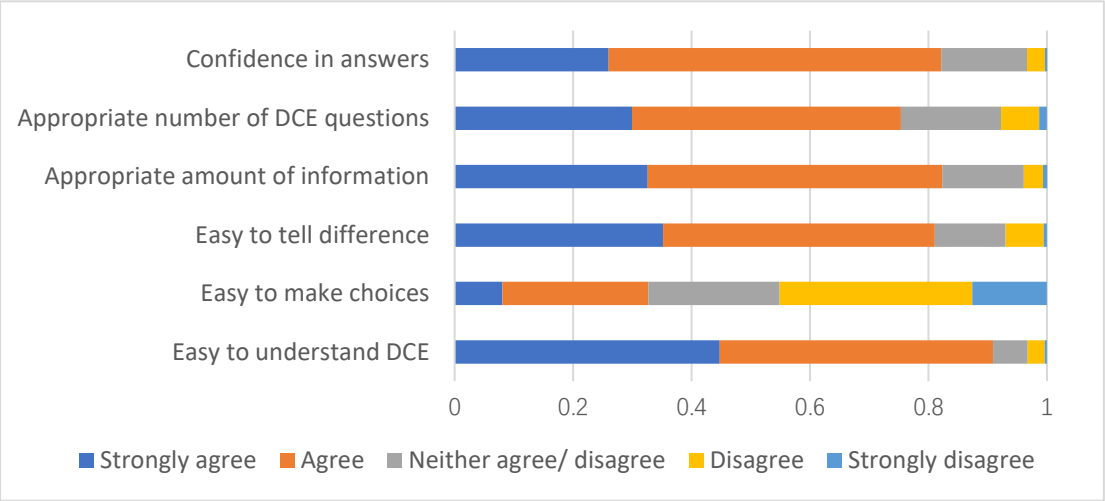
7.2.1 Understanding and data quality check with UK sample

The data collection received 26039 valid DCE answers from UK sample. The standard-order DCE_{TTO}, revised-order DCE_{TTO} and the DCE-Death received 14, 13 and 17 responses for each DCE profile, including the dominated pairs.

The average time to complete the survey was 17 minutes, with shorter time taken for DCE_{TTO} (16.5 min) and longer for the DCE with death survey versions (18.7 min). The average time for each DCE question was 34 seconds, ranging from 28 seconds (DCE_{TTO}) to 31 seconds (DCE with death). 27% of respondents spent less than 10 minutes to complete the survey, while 15% of respondents spent less than 10 seconds for at least one DCE question (Table 3, Appendix E). The average time for the first three and last three questions did not have significant difference for all cohorts, indicating the average DCE response time revealed the general situation.

Most respondents felt confident with their choices (82%), agreed that the number of choices sets per person was appropriate (75%) and the amount of information provided was appropriate (82%). 81% of respondents agreed that the DCE choices were easy to tell the difference between the given health states (Figure 13).

Figure 13 UK participant feedback



Participants were able to provide free text feedback at the end of the survey, where 4% of all respondents provided negative feedback or design suggestions for this survey in each group. Around 3% of respondents reported an inconsistent answer for the repeated employment status question. 1% of respondents always selected the left or right health state for all of the questions. However, this left-right choice did not provide information about whether data had position bias, or the health state randomization always placed the better alternative at a certain spot[315, 316]. 31% of all respondents self-reported considering all of the information presented all the time, which was the ideal decision-making strategy for DCE research. The other most employed decision-making strategies were considering only the subset of HWB attributes that the participant believed to be important (36%), considering different HWB attributes (16%) and mainly consider the length of time (12%). DCE_{TTO} (Design 1 and 2) had higher proportion of considering length of time in the decision than the DCE-Death. Among those who never chose a health state as WTD, 11% concluded that the reason was that the presented health states were not “bad enough” to be worse than death. 5% stated that it was difficult to imagine “immediate death” (Table 17).

In conclusion, the data quality check found that Design 1 and 2 had more speeder, whereas Design 3 had more left-right heuristic respondents. A larger proportion of respondents considered all of the information in making DCE choice in Design 3.

Table 17 Data quality check and WTD reasoning with UK data

		Design 1 (n=627)		Design 2 (n=600)		Design 3 (n=776)		Overall (n=2003)	
		N	%	N	%	N	%	N	%
Data quality check									
Negative feedback		27	4.31%	24	4.00%	32	4.12%	83	4.14%
Completed survey in less than 10 minutes		196	31.26%	177	29.50%	160	20.62%	533	26.61%
Completed DCE task in less than 10 seconds[1]		100	15.95%	99	16.50%	109	14.05%	308	15.38%
Repeated employment status question inconsistency[2]		17	2.71%	12	2.00%	26	3.35%	55	2.75%
Left-right bias: left		3	0.48%	4	0.67%	11	1.42%	18	0.90%
Left-right bias: right*		1	0.16%	1	0.17%	6	0.77%	8	0.40%
Decision-making strategies (self-report)[3]									
Considered all of the health and wellbeing aspects all the time		173	27.59%	166	27.67%	287	36.98%	626	31.25%
Only considered health and wellbeing aspects that I believe to be important		229	36.52%	201	33.50%	293	37.76%	723	36.10%
Considered different health and wellbeing aspects each time		98	15.63%	87	14.50%	139	17.91%	324	16.18%
Mainly considered the length of time in Life A or B		106	16.91%	120	20.00%	24	3.09%	250	12.48%
Considered health and wellbeing aspects not presented here		2	0.32%	1	0.17%	5	0.64%	8	0.40%
Other decision method		11	1.75%	16	2.67%	15	1.93%	42	2.10%
Select randomly		2	0.32%	3	0.50%	3	0.39%	8	0.40%
Do not know		6	0.96%	6	1.00%	10	1.29%	22	1.10%
Participants who never selected WTD						79	11.28%		
Reason for not selecting WTD state (Cohort 3) [2]									
	In the questions, there were always better options in either Life A or Life B.					60	7.73%		
	Being alive, even with the given health and wellbeing problems, is always better than being dead.					17	2.4%		
	I choose “immediate death” as worst because of my religious beliefs.					6	0.77%		
	I choose “immediate death” as worst because of my outlook on life or family related considerations.					48	6.19%		

	I found it difficult to imagine what “immediate death” would mean so I did not consider it.	9	1.16%		
Reason for not selecting WTD state – because of study design only[4]		18	10.72%		
Reason for not selecting WTD state – because of dead state hard to imagine only		4	4.92%		

Notes: * The Design 3 group counted respondents always selecting the middle health state

[1] completion time under 10 seconds with at least one DCE task

[2]This number is after the check of choice understanding

[3]more than 1 decision-making strategy can be selected

[4]the presented health states in all of the choice sets were not bad enough to be compared with dead state

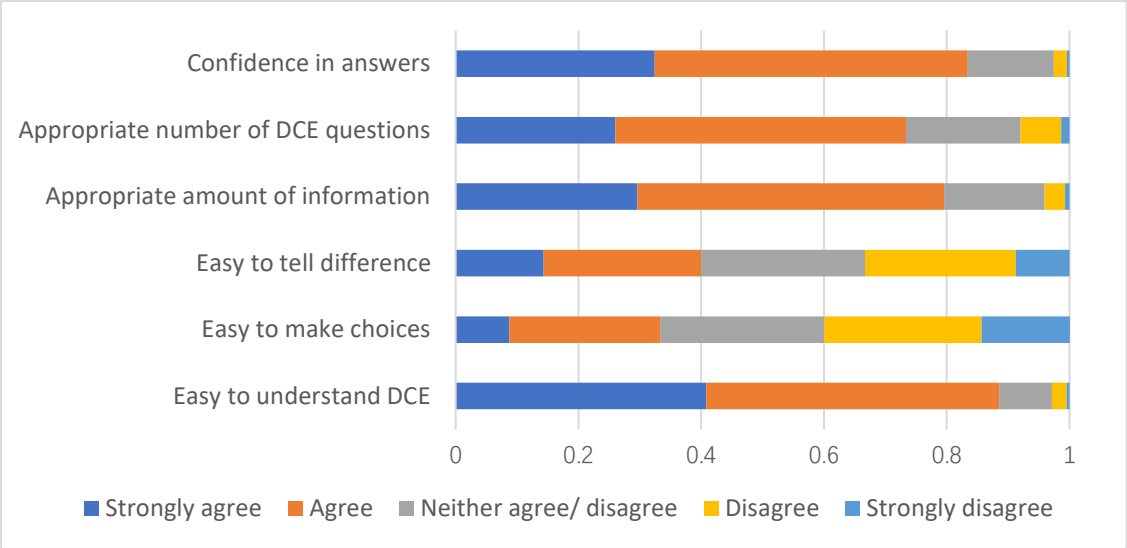
7.2.2 Understanding and data quality check with Australian sample

The data collection received 26039 valid DCE answers from Australian sample. The standard-order DCE_{TTO}, revised-order DCE_{TTO} and the DCE-Death received 14, 13 and 17 responses for each profile, including the dominated pairs.

An average time spent was 19 minutes to complete the whole survey, with shorter time taken for DCE_{TTO} (18.7 min) and longer for the DCE with death (20.2 min). The average time to complete each DCE question was 29 seconds, ranging from 28 seconds (DCE_{TTO}) to 30 seconds (DCE with death). 21% of respondents spent less than 10 minutes to complete the survey, while 19% of respondents spent less than 10 seconds for at least one DCE question (Table 4, Appendix E). 61% of all speeder records appeared with the last three questions, indicating that the last questions might have higher risk of heuristic decision.

The majority of respondents reported the DCE questions as easy to understand (96%), confident in answers (84%), the number of DCE questions was appropriate (83%) and had appropriate amount of information to make their decisions (80%). However, less than half of respondents found it easy to tell the difference or make the choice (33% and 40%), which were both lower than the correspondent rates with UK sample. More Design 3 respondents disagreed with making choices easily, but had slightly higher percentage agreed they were confident in answers (Figure 14).

Figure 14 Australian participant feedback



In the optional free text feedback at the end of the survey, 3% of all respondents provided negative feedback or design suggestions for this survey, with limited variations among the designs. 4% of respondents reported an inconsistent answer for the repeated employment status question. 3% of respondents always selected the left or right health state for all of the DCE questions, where Design 3 had a higher proportion. The 37% of all respondents considered all of the information presented in the DCE task all the time, with a higher proportion with the Design 3 (42%) in comparison to Design 1 and 2. 12% of all non-WTD respondents explained the reason as the presented health states were not “bad enough” to be worse than death, while 3% stated that it was difficult to imagine “immediate death” (Table 18).

The Australian sample had higher proportion of left-right bias compared with the UK sample, but the proportion of respondents considering all of the information was higher as well. However, for the Design 3, the proportion of respondents selecting no health state as WTD doubled, whist the number of non-WTDs stated out of the DCE design and imagination difficulties remained similar (28 versus 22 for Australian and UK samples). This comparison indicated that UK and Australian samples had similar data quality, but had some difference on decision-making due to culture or religious reasons.

Table 18 Data quality check and WTD reasoning with Australian data

		Design 1		Design 2		Design 3		Overall	
		N	%	N	%	N	%	N	%
		632	100.00%	604	100.00%	783	100.00%	2019	100.00%
Data quality check									
Negative feedback		15	2.37%	12	1.99%	29	3.70%	56	2.77%
Completed survey in less than 10 minutes		146	23.10%	121	20.03%	161	20.56%	428	21.20%
Completed DCE task in less than 10 seconds[1]		129	20.41%	108	17.88%	145	18.52%	382	18.92%
Repeated employment status question inconsistency[2]		16	2.53%	25	4.14%	30	3.83%	71	3.52%
Left-right bias: left		12	1.90%	9	1.49%	31	3.96%	52	2.58%
Left-right bias: right*		1	0.16%	4	0.66%	3	0.38%	8	0.40%
Decision-making strategies (self-report)[3]									
Considered all of the health and wellbeing aspects all the time		216	34.18%	193	31.95%	328	41.89%	737	36.50%
Only considered health and wellbeing aspects that I believe to be important		192	30.38%	201	33.28%	274	34.99%	667	33.04%
Considered different health and wellbeing aspects each time		85	13.45%	90	14.90%	130	16.60%	305	15.11%
Mainly considered the length of time in Life A or B		91	14.40%	81	13.41%	15	1.92%	187	9.26%
Considered health and wellbeing aspects not presented here		9	1.42%	11	1.82%	9	1.15%	29	1.44%
Other decision method		19	3.01%	15	2.48%	16	2.04%	50	2.48%
Select randomly		8	1.27%	6	0.99%	4	0.51%	18	0.89%
Participants who never selected WTD						186	23.75%		
Reason for not selecting WTD state (Cohort 3) [2]									
	In the questions, there were always better options in either Life A or Life B.					64	8.17%		
	Being alive, even with the given health and wellbeing problems, is always better than being dead.					135	17.24%		
	I choose “immediate death” as worst because of my religious beliefs.					13	1.66%		
	I choose “immediate death” as worst because of my outlook on life or family related considerations.					44	5.62%		
	I found it difficult to imagine what “immediate death” would mean so I did not consider it.					24	3.07%		

Reason for not selecting WTD state – because of study design only[4]	24	3.07%		
Reason for not selecting WTD state – because of dead state hard to imagine only	4	0.51%		

Notes: * The Design 3 group counted respondents always selecting the middle health state

[1] completion time under 10 seconds with at least one DCE task

[2]This number is after the check of choice understanding

[3]more than 1 decision-making strategy can be selected

[4]the presented health states in all of the choice sets were not bad enough to be compared with dead state

7.3 Regression outcome

7.3.1 Model recap

The DCE_{TTO} data was analysed using a conditional logit model. Duration was modelled as a linear variable, with an assumption that respondents follow linear time preference. The coefficients are anchored onto the 1 to 0 full health-dead scale by dividing each attribute level coefficient with the duration level coefficient.

The DCE-Death data was analysed using a model with no time interaction (time does not differ across the DCE profiles), with conditional logit model and the rank-order logit model. The coefficients were anchored onto the 1 to 0 full health-dead scale by dividing each attribute level coefficient with the dead dummy coefficient.

Models 1, 2 and 3 are conditional logit regression with the three cohorts: Model 1 is DCE_{TTO} with the health attributes presented first in the task; Model 2 is DCE_{TTO} with the health attributes presented first in the task; Model 3 is DCE-Death. Model 4 is the rank-order logit regression with DCE-Death data. A pooled model also conducted with all of the DCE_{TTO} data to test the significance of information order on preference. This modelling result is presented in Section 7.3.3. More information about the regression models, including function and coding, see the Section 6.8.2.

7.3.2 Model performance

Model 1 – Model 4 were listed in Table 19, where all of the models were significant with a 0.05 significance level. By using the conditional logit model (Model 3) and the rank-order logit model (Model 4) with the same sample data, they produced similar range of utility and the distribution of utility values, but rank-order logit models had smaller Pseudo R-squared statistic (Table 19). This study used Model 3 to generate the DCE-Death health state valuation result. The Pseudo R-squared value for Models 1-3 were between 0.097 to 0.191, within the range of values that other DCE valuation studies reported[317, 318].

Duration coefficients in Model 1 and 2 with UK and Australia samples were significant and had the expected positive sign. DCE-Death data regression followed the RUM was expected to be negative, with a negative sign indicating the distance from full health to death. Dead dummy in Model 3 and 4 with UK and Australia samples were significant and had negative sign for disutility (Table 5 and Table 6, Appendix E, with UK and Australian sample).

Table 19 Summary of key findings from the models

Country		Model 1	Model 2	Model 3	Model 4
UK	Log-likelihood statistics[1]	4544.945	4393.045	18443.209	15970.090
	Prob > chi2[2]	<0.001	<0.001	<0.001	<0.001
	Pseudo R-squared[3]	0.191	0.181	0.121	0.071
	Logical inconsistent	10	14	4	4
	Insignificant 10%	17	19	17	17
	Utility of health state with all items at level 1	1	1	1	1
	Utility of health state with all items at level 2	0.686	0.822	0.764	0.758
	Utility of health state with all items at level 3	0.448	0.556	0.453	0.452
	Utility of health state with all items at level 4	-0.179	-0.083	-0.133	-0.13
	Utility of health state with all items at level 5	-0.634	-0.791	-0.887	-0.862
	Scale length[4]	1.634	1.791	1.887	1.862
	Mid-point to length[5]	33.8%	24.8%	29.0%	29.4%
	Proportion of WTD	17.16%	16.74%	27.40%	27.36%
AUS	Log-likelihood statistics[1]	4919.716	4640.015	18772.880	16420.360
	Prob > chi2[2]	<0.001	<0.001	<0.001	<0.001
	Pseudo R-squared[3]	0.139	0.143	0.097	0.061
	Logical inconsistent	8	11	4	2
	Insignificant 10%	26	15	11	14
	Utility of health state with all items at level 1	1	1	1	1
	Utility of health state with all items at level 2	0.699	0.598	0.673	0.711
	Utility of health state with all items at level 3	0.599	0.488	0.472	0.483
	Utility of health state with all items at level 4	-0.089	-0.025	-0.146	-0.126
	Utility of health state with all items at level 5	-0.588	-0.539	-0.713	-0.699
	Scale length[4]	1.588	1.539	1.713	1.699
	Mid-point to length[5]	25.3%	33.3%	30.8%	30.4%
	Proportion of WTD	13.16%	12.75%	17.01%	17.05%

Notes: [1]The Log likelihood statistics is the value of final model. This value cannot be directly compared between models as they used different survey data.

[2]The model significance is evaluated through chi-square statistics. The model is significant if the value is less than 0.05

[3]Pseudo R-squared summarizes the proportion of variance explained by the independent variables. A larger R-squared statistic indicates better explanatory power. This value cannot be directly compared between models as they used different survey data and model function.

[4]Scale length is the difference between utility values for states with level 1 for all dimensions and level 5 for all dimensions

[5]Midpoint to length is the value assessed by dividing the difference between utility value for states with level 1 for all dimensions and level 3 for all dimensions by the scale length.

Table 20 Non-significant levels, positive disutility and non-monotonic levels

		UK			AUS		
		Non-significant[1]	**Non-monotonic & incorrect sign[1]**	**Significant & non-monotonic[1]**	**Non-significance[1]**	**Non-monotonic & incorrect sign[1]**	**Significant & non-monotonic[1]**
By Attributes							
Vision		0	0	0	1	0	0
Hearing		2	0	0	1	0	0
Mobility		2	1	0	2	1	1
Daily activity		2	0	0	1	1	0
Sleep		8	5	2	8	6	1
Fatigue		5	4	0	6	1	1
Loneliness		7	6	1	7	3	0
Support		5	2	1	3	4	2
Concentrating		3	3	0	8	4	0
Control		7	4	0	5	3	0
Anxious		9	6	0	6	4	1
Sad/depress		4	3	0	5	2	0
Pain severity		1	0	0	0	0	0
By Levels							
Level 1-2	Severity[2]	6	0	0	2	0	0
	Frequency[3]	17	10	1	18	8	0
Level 2-3	Severity[2]	1	1	0	3	2	1
	Frequency[3]	18	10	0	18	10	1

Level 3-4	Severity[2]	0	0	0	0	0	0
	Frequency[3]	10	8	0	6	4	1
Level 4-5	Severity[2]	0	0	0	0	0	0
	Frequency[3]	3	5	3	6	5	3

Notes: **Highlighted in bold and italics**: attributes and levels that has the largest number of non-significant levels, non-monotonic coefficients and incorrect sign with both countries. The disordered & incorrect sign column signifies that the regression coefficient has a positive sign for the disutility, or the disutility of worse level is smaller than the disutility of adjacent better level. The significant & non-monotonic are the significant regression coefficients that has a positive sign for the disutility, or the disutility magnitude of a lower level is smaller than that of a higher level.

Non-significance level: 10%.

[1]By attribute: sum of the number of non-significant levels, positive disutility and non-monotonic coefficients in three models. The total number of coefficients for each attribute is 12 since this is summing the results for Models 1, 2 and 3. The total number of coefficients for each level is 39.

[2]Severity levels: HWB dimension attributes with health problems described by its severity (No difficulty, Slight difficulty, Some difficulty, A lot of difficulty, Unable). Attributes are Vision, Hearing, Mobility, Daily activity and Pain severity. This is summing the results for Models 1, 2 and 3.

[3]Frequency levels: HWB dimension attributes with health problems described by its frequency (None of the time, Only occasionally, Sometimes, Often, Most or all of the time). Attributes are Sleep, Fatigue, Loneliness, Support, Concentrating, Anxious, Sad/depress and Control. This is summing the results for Models 1, 2 and 3.

See Table 7, Appendix E for more information.

The utility ranges were from 1 to -0.634 (Model 1) or -0.887 (Model 3) with the UK sample, and 1 to -0.588 (Model 2) and 1 to -0.713 (Model 3) with the Australian sample (Table 19).

There was some evidence of non-significant levels, positive disutility and non-monotonic levels with all designs with both UK and Australian data, especially with wellbeing attributes and level 2/3 coefficients. Wellbeing attributes, such as sleep, anxious, loneliness, support (Australian data) and control (UK data) were the attributes with the largest number of non-significant and non-monotonic/ incorrect sign levels (Table 20). The frequency levels had more insignificance and logical inconsistency than the severity levels in all of the level effects. The Model 1 regression with Australian data had the largest number of non-significant and non-monotonic levels, followed by the Model 2 regression with UK data (Table 20).

Level 2 and level 3 effects were more likely to be insignificant and non-monotonic. Attributes Sleep, support and the attribute level 5 (with all attributes) had the largest number of significant coefficients with non-monotonic disutility, but there was no significant coefficient with non-monotonic disutility. All of the significant coefficient with non-monotonic disutility appears in Model 1 and 2 (Table 7, Appendix E). In summary, Model 3, physical and mental health attributes and severity levels performed better than Models 1 and 2, wellbeing attributes and frequency levels, from the perspective of significance, monotonicity and difference distinguishment.

This study asked respondents to select the five most important attributes after completing all of the DCE choice sets (stated ranking). The regressed rank used the anchored magnitude of the worst-level disutility as an indicator of the relative importance of each model. UK and Australian participants ranked Pain severity, Vision, Daily activity, Mobility and Hearing as the most important attributes affecting their DCE decisions, where the regression analysis with Australian outcome generated same result (Figure 15, or Table 5 and Table 6, Appendix E). UK Model 1 regression outcome ranked Anxious, instead of Hearing, as the five most important

attributes. The attributes with the largest level-five disutility with the three models were always Pain severity and Vision. 5 out of the 6 most highly valued attributes (i.e. with the largest utility decrements) were health attributes, with 3 attributes similar to EQ-5D dimensions (Pain severity – pain/discomfort; Daily activity - usual activity; Anxious - anxiety/depression)[319].

The stated and regressed preference on attributes show high consistency. Figure 15 matched the two rankings with each design. A higher proportion of attributes fell near to the line Y=X indicating a higher stated and regressed ranking consistency. A higher proportion of attributes ranked top-five fell on the Y=X line, while the other attributes showed more variance. DCE-Death design data revealed a higher proportion of attributes near to the line.

Figure 15 Rank order of stated and regressed preference by models (from top to bottom: Model 1, Model 2, Model 3) and by country (from left to right: UK and Australian datasets)

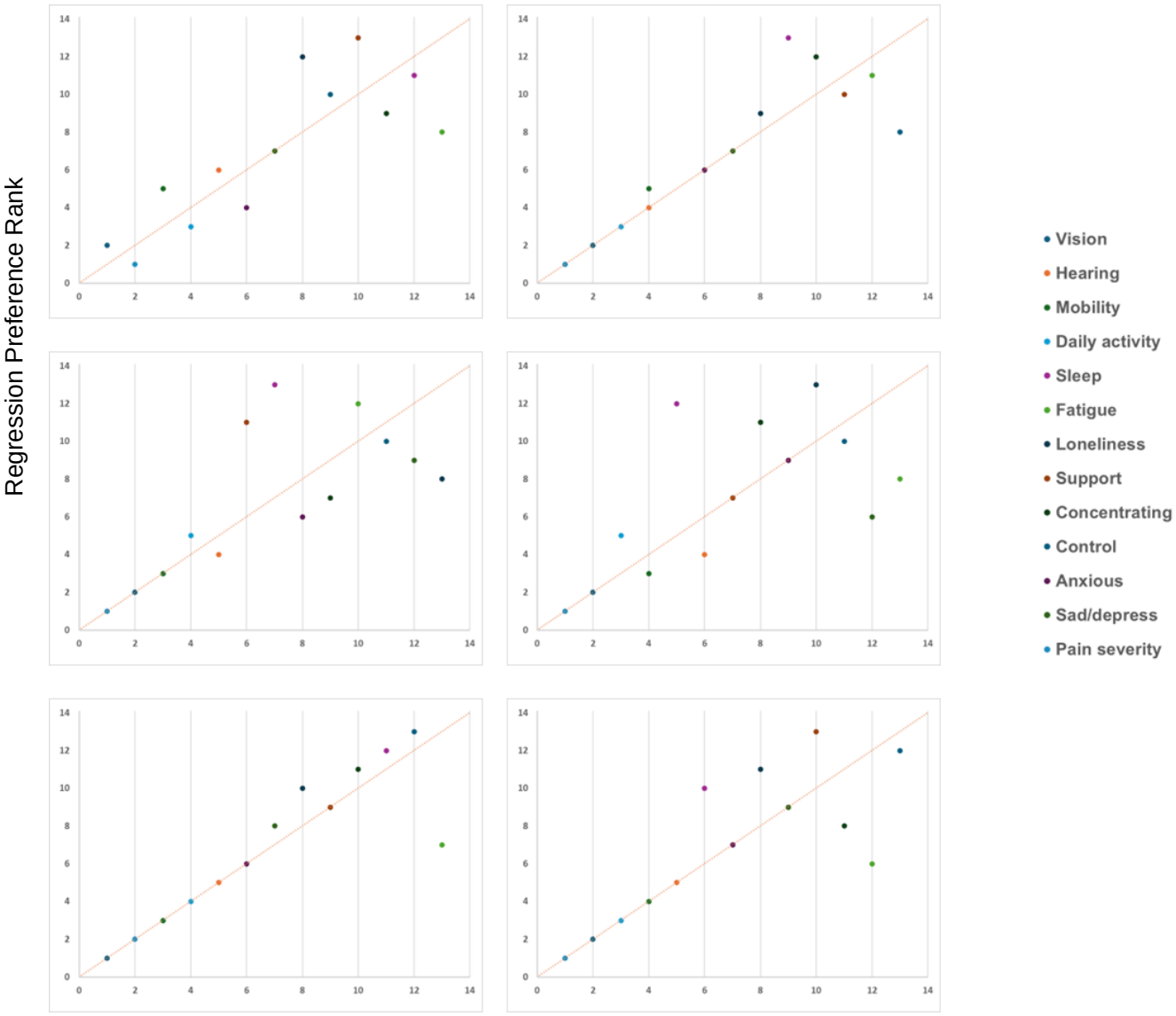
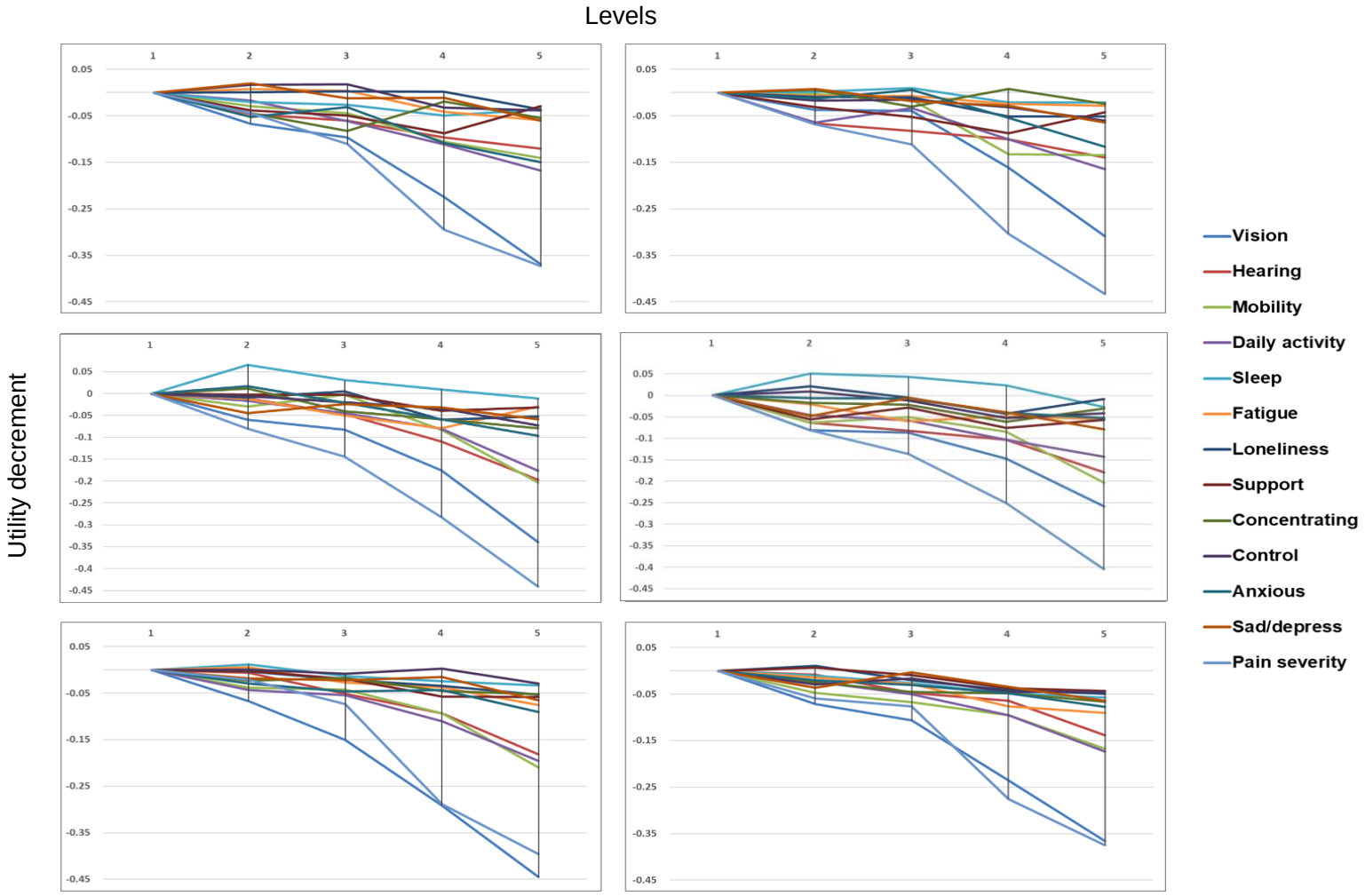


Figure 16 reports the marginal effect of moving from one level to another, by attributes. A value above 0 indicated a positive disutility. The marginal effects, which was represented by the slopes of the graphs, have steeper utility descents for level 3 – level 4 and level 4 - level 5. Physical health attributes, for example, Vision, Pain severity, Daily activity, Mobility and Hearing, had a “kink” moving from level 3 – level 4. Other attributes’ decrements were more linear. Physical health attributes with severity levels (Figure 1a, Appendix E) had larger variance, and wellbeing/mental health attributes (Figure 1b, Appendix E) with frequency levels was centered on disutility range 0 to -0.1.

All models had high robustness that results did not show significant variance with sub-group passed dominance questions and reported high understanding.

Figure 16 UK (left) and Australian (right) disutility value by level and dimensions, by models (from top to bottom: Model 1, Model 2, Model 3)



7.3.3 Order effect

The utility ranges are 1 to -0.634/-0.588 (Model 1/Model 2) and 1 to -0.791/-0.539 (Model 1/Model 2) with UK and Australian samples respectively. With the UK responses, Model 2 mild health states utility values (from perfect health to state 4444444444444) are higher than of the Model 1, while extreme states (from state 4444444444444 to state 5555555555555) are lower. With the Australian responses, the Model 2 mild health states utility values (from state 2222222222222 to state 333333333333) are lower than that of the Model 1, while extreme states (from state 4444444444444 to state 5555555555555) are higher (Table 19). However, it should be noted that the magnitude of utility difference is small.

The Model 2 with both UK and Australian samples had more positive or non-monotonic samples. On the other hand, there was mixed evidence on the number of insignificant level coefficients: the UK sample had a larger number of insignificant levels with Model 2 and Australian sample with Model 1. By comparing the number of non-monotonic and significant attributes, Model 2 had a larger number with both UK (1 versus 3) and Australian (1 versus 5) samples (Table 7, Appendix E).

Only 5 out of 78 coefficients were significantly different with a significance level of 5% by conducting Wald test, with UK and Australian data (Table 21). However, the preference with single level coefficient might not reflect the overall effect on relative preference of health and wellbeing attributes. By calculating the relative importance of physical/mental health attributes (Vision, Hearing, Mobility, Daily activity, Sleep, Fatigue, Sleep, Sad/depression, Pain severity) versus the wellbeing attributes (Fatigue, Loneliness, Support, Concentrating, Control, Anxious), it yielded a global evaluation of preference variation. The relative importance value did not alter with UK sample, but the number increased from 4.4 to 7.1 with Australian value set (Table 22).

Table 21 Model 1 V.S. Model 2 coefficient Wald test

	Australian data			UK data		
	Level 3	**Level 4**	**Level 5**	**Level 3**	**Level 4**	**Level 5**
Seeing	2.71	0.17	2.31	0.10	2.15	0.34
Hearing	1.45	0	1.51	0.01	1.38	9.42
Getting around	1.09	1.93	4.20	2.15	0.38	6.78
Day to day activities	0.62	0.03	0.41	0.22	0.65	0.73
Sleeping	0.74	1.56	0.04	4.22	4.47	0.69
Exhausted	2.17	0.58	0.65	4.08	1.74	1.05
Lonely	0.03	0.03	1.35	0.02	3.65	0.29
Unsupported	0.38	0.12	0.21	2.80	2.74	0
Thinking	0.08	4.11	0.02	0.50	3.07	1.42
Anxious	0.02	0.32	0.37	0.71	0.20	0.71
Depression	0.10	0.17	3.46	0.28	2.08	1.74
Control	0.14	0.15	0.22	0.2	0.42	0.01
Physical pain	0.37	2.65	0.77	0.71	0.51	3.38

Notes: Values given here are Wald test score; the p value over 0.1 and 0.05 are highlighted with dark or light green. Dark green: p value > 0.1; light green: 0.1> p value > 0.05

Australian data: comparison of results regressing with the Australian sample.

UK data: comparison of results regressing with the UK sample.

Secondly, attribute order influenced the utility distribution, with varied influence on mild (from state 2222222222222 to state 333333333333) or extreme states (from state 4444444444444 to state 5555555555555). However, it should be noted that the magnitude of utility difference was small.

Table 22 relative importance[1] of physical/mental health attributes versus the wellbeing attributes

	UK	Australia	Absolute difference
Model 1	4.314	4.376	0.062
Model 2	4.040	7.098	3.058
Model 3	5.632	5.079	0.353
Average (Model 1 and 3)	4.971	4.721	0.250
Absolute difference (Model 1 and Model 2)	0.266	2.722	
Absolute difference (Model 1 and Model 3)	1.318	0.703	

Notes: [1]Anchored dividing sum of level 5 disutility of Vision, Hearing, Mobility, Daily activity, Sleep, Fatigue, Sad/depression, Pain severity, by the sum of level 5 disutility of Sleep, Fatigue, Loneliness, Support, Concentrating, Control, Anxious

The conditional logit model allowing for the order effect interactions with each term was conducted. Choice data collected with Design 1 and Design 2 was pooled for single regression. Table 5 and Table 6, Appendix E recorded the significant interaction of each level dummy with the order dummy (standard order DCE_{TTO} – 0, revised-order DCE_{TTO} – 1), and the order dummy with duration interaction. With DCE_{TTO} data collected in the UK, the attribute order effect on time was not significant, though order effects with Sleep (level 2 and level 4) and Depression (level 2) levels were significantly positive, and with Hearing (level 5), Getting around inside and outside (level 5), Control (level 2) levels were significantly negative. With Australian sample, the attribute order effect on time was significantly positive, indicating duration weighted more in the revised-order design. Hearing (level 5), Getting around inside and outside (level 2 and level 5), Concentration (level 4) levels had a significant negative order effect. A full pooled regression with all interactive terms is presented in Appendix E and the significant interactive terms is reported above.

In conclusion, the order effect analysis supported the assumption that the order of health and wellbeing attributes influenced the respondent preference with some of the attributes but had mixed evidence in two countries. the survey data revealed

opposite utility distribution influence and midpoint to length in the two countries. The proportions of health states classified as WTD were close (around 16% for UK and 13% for Australian sample), but Model 2 proportion was always lower in both countries. The pooled regression outcomes in both counties revealed that by listing the wellbeing attributes first, respondents had higher disutility for physical health attributes hearing and getting around but had conflict evidence for other HWB attributes and duration.

7.3.4 Design effect

Model 1 had a larger number of non-monotonic/positive disutility and insignificant attribute level coefficients than Model 3 with Australian data sets, while had the same insignificance performance with UK sample as Model 3 had a larger number of significant and monotonic wellbeing levels (Table 19). There were no Model 3 anchored coefficients that were significant and non-monotonic, while the number was 1 with Model 1 in each country. With both designs, the respondents ranked Seeing, Mobility, Daily activity and Physical pain as the most important attributes. Model 3 achieved more alignment between the stated preference and regressed ranking comparison (Figure 5).

Model 1 and Model 3 showed considerable difference in terms of scale length and utility distribution. The difference of scale length could be interpreted as a willingness indicator to improve QoL (Model 1) by sacrificing time, or to prevent death (Model 3) by enjoying less health. The utility range with Australian sample was 1 to -0.588 and -0.713 for Models 1 and 3, and the Model 3 utility values were always smaller than the Model 1 values, indicating a higher willingness to trade-off. Model 3 utility value of health states 2222222222222, 3333333333333 and 4444444444444 were slightly higher than the corresponding Model 1 utilities with UK sample, but Model 1 was higher with the Australian sample. On the other hand, the Model 3 utility for the worst state (5555555555555) was lower than Model 1 utility, with both UK and Australian sample.

By calculating five million Monte Carlo simulations for each group, the Model 3 WTD proportion was lower than the Model 1 in both UK and Australia (Table 19), which was consistent with the finding that Model 3 had a lower worst state.

The impact of different DCE designs on attribute level preferences was analyzed. All level 5 and level 4 anchored coefficients, except for Concentrating/Thinking Clearly, showed statistically significant differences across both the UK and Australian samples (Table 23). Beyond level preference comparisons, we conducted paired

comparisons of health state values across the design methods. The mean DCE_{TTO} value (0.32 and 0.26 with UK and Australia value set) in both countries were higher than the corresponding values calculated by the DCE_{death} value set (0.26 and 0.24 with UK and Australia value set). Appendix E presents the results ranked by DCETTO values, indicating that DCE_{death} values were generally lower, with greater variance observed in the Australian sample.

Despite these differences, the Pearson's correlation coefficient reveals a strong correlation (>0.9) between state values derived from each design, suggesting consistency in respondents' preferences. Additionally, the intraclass correlation coefficient (ICC) was below 0.05, indicating that respondents were likely to make similar DCE choices regardless of design (Appendix E).

Table 23 Model 1 V.S. Model 3 coefficient Wald test

	Australian data			**UK data**		
	Level 3	**Level 4**	**Level 5**	**Level 3**	**Level 4**	**Level 5**
Seeing	33.81	224.76	16.51	125.58	450.64	827.52
Hearing	7.76	30.52	86.99	19.69	49.02	130.43
Getting around	11.3	74.49	128.81	16.99	81.55	235.5
Day to day activities	7.85	51.68	145.95	25.49	94.88	224.2
Sleeping	0.49	6.88	9.22	3.22	10.85	9.7
Exhausted	1.69	11.44	18.55	0.56	9.17	35.52
Lonely	2.42	10.86	12.57	1.02	3.39	14.32
Unsupported	2.6	14.38	8.45	3.87	27.2	12.43
Thinking	8.01	3.06	14.24	14.47	7.49	23.16
Anxious	1.89	7.85	16.28	0	1.71	10.76
Depression	0.67	11.94	41.17	7.24	33.65	83.27
Control	0.67	4.54	20.68	1.81	1.81	27.08
Physical pain	51.16	356.22	531.3	70.43	512.52	728.21

Notes: Values given here are Wald test score; P value is the probability of obtaining the Wald test statistic given that the null hypothesis is true, which is compared to the critical value 0.05 and 0.1 to determine if the difference is significant. The difference is significant if the P value is smaller than the critical value; the p value over 0.1 and 0.05 are highlighted with dark or light green. Dark green: p value > 0.1; light green: 0.1> p value > 0.05

Australian data: comparison of results regressing with the Australian sample.

UK data: comparison of results regressing with the UK sample.

Separate regressions were conducted to generate DCE_{death} estimates using only respondents who identified at least one health state as equal to or WTD, thus contributing to the anchoring of health states on the QALY scale. In both the UK and Australian samples, the utility values for the worst health states were lower (-0.89 and -1.15, respectively), and model fit improved, as all respondents in this subset acknowledged the existence of WTD states.

When comparing the utility values for selected health states from this subset with those generated by the DCETTO sample, a larger proportion of health states showed lower utility values in the DCE_{death} group. However, Pearson's correlation coefficient and ICC results remained consistent (see Supplementary Material Part B). Thus, excluding non-WTD selectors produced lower utility values but did not alter the overall similarity in utility value distributions between the DCE_{TTO} and DCE_{death} designs.

In conclusion, attribute level coefficient Wald test and utility distribution proved that the health state preferences systematically differed between the Model 1 and Model 3. Significant effects were observed in the Wald test of level 4 and level 5, number of significant and monotonic attribute levels, utility distribution and utility range. Model 3 had smaller proportions of the significant and non-monotonic anchored value, wider range of utility and higher WTD proportion. However, this WTD difference should be interpreted noting the amount of information used for setting WTD by the two designs: Model 1 with the DCE_{TTO} design collected anchoring information by introducing the duration trade-off to all participants and with all questions, while the Model 3 with the DCE-Death design gathered anchoring information from participants "believe" the WTD state existed. A paired comparison with the selected state values indicates that there were significant point estimation differences. However, the values still varied in a similar way.

7.3.5 UK and Australian preference difference

The comparison of preference in two countries included the examination of the ranking and relative importance of dimensions, relative decrements between levels, scale length difference and the distribution of utility, with Model 1 and Model 3 representing the DCE_{TTO} and DCE-Death results.

The Table 19 included the midpoint to length that all of the four values were around 30%. The minimum utility values dropped to -0.59/-0.63 and -0.71/-0.89 (Model 1/3) when applying Australia and UK weights, suggesting that other things being equal, the UK sample considered a deficiency in HWB worse than the Australian sample for the more severe states. A larger proportion of health states were classified as WTD with UK sample, despite UK utilities for health states better than state 2222222222222 were higher than Australian utilities, indicating that the utility difference was not linearly changed across all health states.

Both countries ranked Vision, Hearing, Mobility, Pain severity as the most important attributes, with an average physical/mental health versus wellbeing attribute importance weight of 4.97 and 4.73 (Table 22). Moving from one level to the next one down involved different marginal utility in both countries. Respondents' considered the distance (disutility) from "moderate" to "severe" (level 3 to 4) as the largest.

Figure 7 and 8 illustrated the DCE_{TTO} and DCE-Death health state value in each country with the two value sets. In general, the Model 1 and Model 3 health state utility values with the varied country weights moved up and down simultaneously, with some point estimation differences observed. The difference increased for health states in the last quarter (ie, the health states worse than 4444444444444) with both designs.

Secondly, the variance analysis discussed reasons for health state value variance by designs[320]. Intraclass correlation coefficient (ICC) attributed less than 5% (0.03 and <0.01 respectively) of variance to the country-level sample differences and the Pearson's correlation with both designs were high (0.96 and 0.98 respectively),

indicating the health state value from two samples changed almost linear. The predicted differences reached 0.06 and 0.02 for DCE_{TTO} standard order design (Design 1) and DCE-Death design (Design 3) respectively (Table 24). The higher mean and median utility values occurred when the Australian weights were applied (mean: 0.32 and 0.25; median: 0.37 and 0.26; standard deviation: 0.23 and 0.23), compared with UK weights (mean: 0.26 and 0.24; median: 0.27 and 0.24; standard deviation: 0.23 and 0.27). All the value distributions were left-skewed (mean<median), except for the Australian DCE-Death group. The variance analysis supported visualization figures.

Based on the analysis above, some characteristics of preference difference were identified, but more striking similarities were observed across the two countries in terms of the utility distribution. UK and Australia had similar stated and regressed preference for the included HWB attributes, but making a judgement that Australian and UK samples had similar preference of each attribute level should be with caution.

Figure 17 UK and Australia value sets per selected health state (n=244) with DCE_{TTO} standard order design

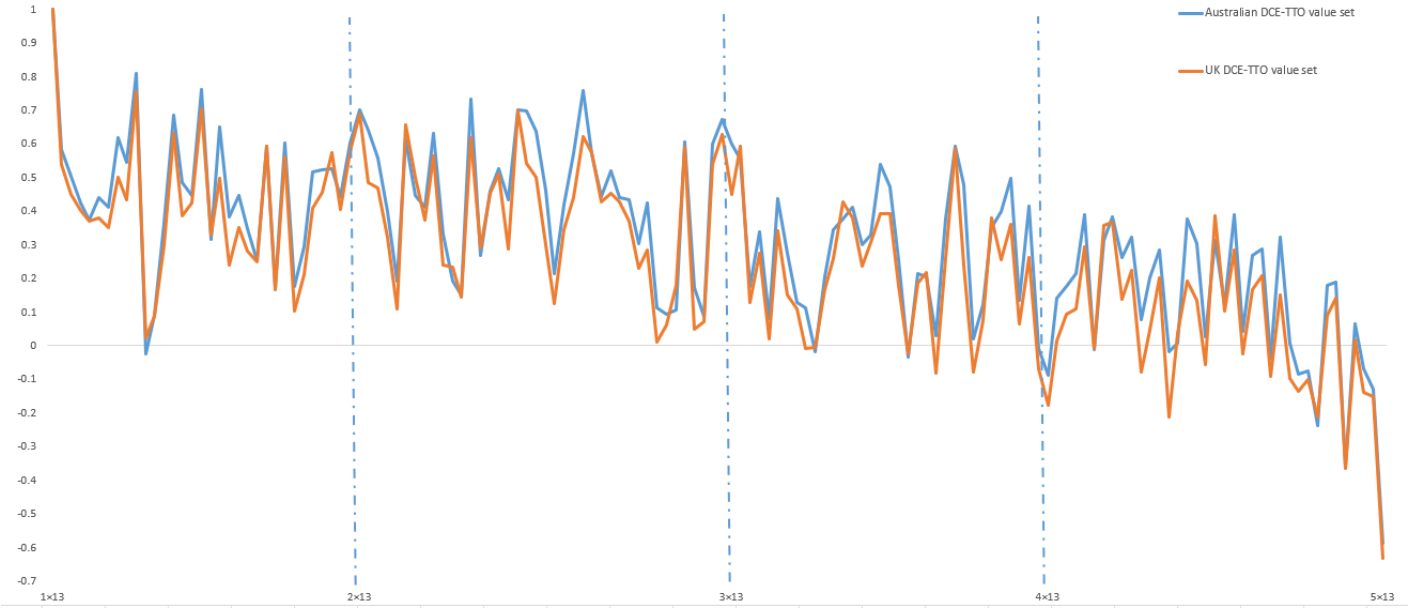


Figure 18 UK and Australia value sets per selected health state (n=244) with DCE-Death design

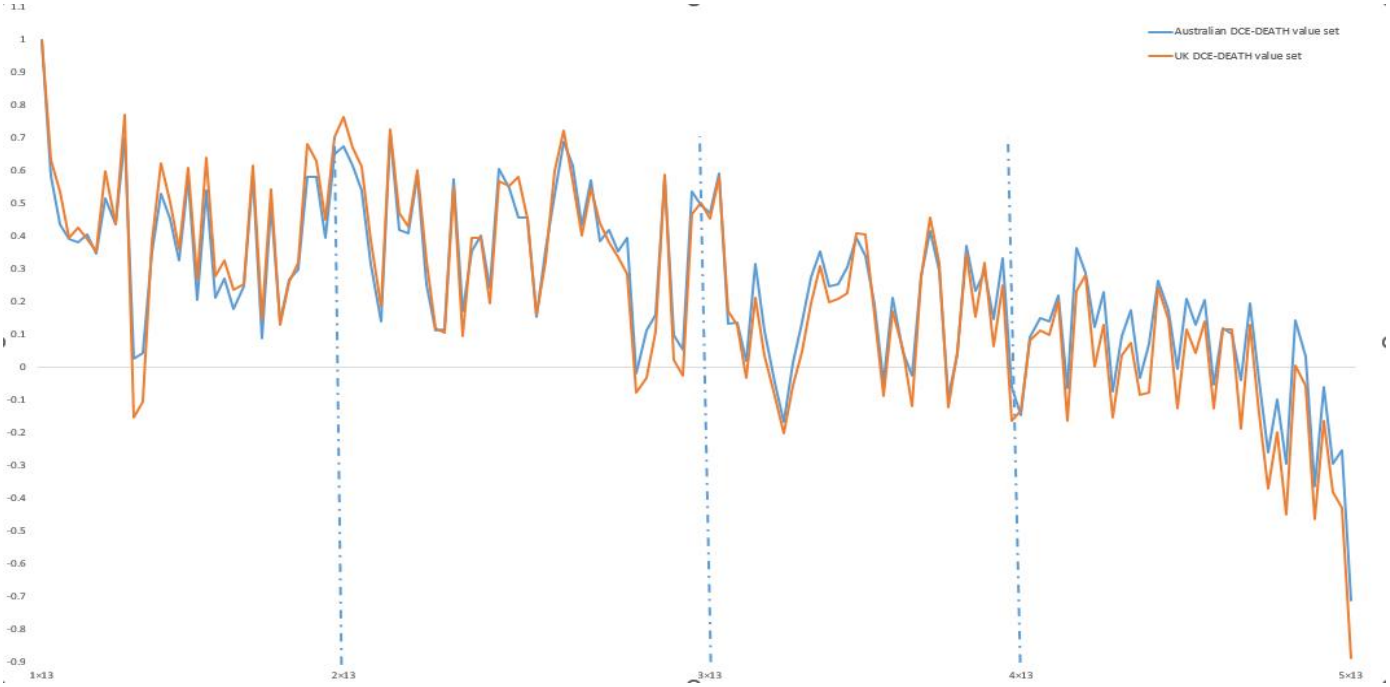


Table 24 correlation and difference analysis with the given health states

		Correlation coefficient analysis		**Pearson's correlation**		**Mean difference**
	Observations	Variance of health state	ICC	Correlation value	T statistics	
DCE_{TTO} design	483	0.932	0.033	0.96	80.44 (P<0.01)	0.06
DCE-Death design	483	0.969	0.003	0.98	113.55 (P<0.01)	0.02

Notes: The correlation analysis generated the average value of 480 efficient design health states in both countries for both designs to analyse the source of health state variance with an ANOVA method, mixed effect model. The 'Variance of health state' was calculated as the variance caused by health state description (row variance/total variance of health state value). The 'ICC' represents the 'intraclass correlation coefficient' to explain the variance due to country difference (column variance/total variance of health state value).

'Pearson's correlation' calculated the linear correlation between data sets in UK and Australia for the model predictions. A higher correlation (between 0.5 to 1) suggested a strong positive correlation.

The 'Mean difference' provides information on the mean utility difference with the two designs.

7.4 Cross-Attribute Level Effect (CALE) estimation

The CALE regression is to explore if a constrained utility function improves the overall significance as well as the individual level performance. A conditional logit regression results with UK and Australia standard order DCE_{TTO} data were reported in Table 25.

All of the models were significant and had a Pseudo R-squared value larger than 0.1. However, the model fitness with DCE-Death data was worse than DCE_{TTO} model, though the DCE_{TTO} regression doubled the sample size. All of the attribute level five effects, as well as the level effects, are significant at the 5% level for both models, which is an artefact of reducing the number of parameters that need to be estimated. The range of utilities are 1 to -0.638/-0.440 (UK and Australian samples, same below) with DCE_{TTO} data, 1 to -0.789/-0.685 with DCE-Death data. All of the models rank Pain severity and Vision as the attributes with lowest disutility. Compared with the additive utility function, the relative importance of HWB attributes remains the same. However, DCE-Death regression has lower utility values with all of the benchmark health states listed.

Although calculated with varied design and respondents, the level 4 anchored coefficients (range 0.601 to 0.675) and level 3 anchored coefficients (range 0.287 to 0.311) are similar across models. The relative preference on levels holds constant for more severe levels when estimated using CALE. However, the level 2 (range 0.125 to 0.232) coefficients have a variance of 0.113, over 50% of the lowest value.

Compared with additive function, CALE has lower worst state utility value and more skewed distribution. The disutility of the worst state is larger than that of the additive model, especially with the DCE-Death design. Second, the CALE utility value distribution is more left-skewed, with shorter tail and lower kurtosis, indicating that the utility estimations for mild health states (better than 2222222222222) are underestimated but estimations for worse health states (worse than 3333333333333) are overestimated (compared with additive model). The CALE model does not propose a linear transformation of all of the health states but mitigates the utility value

differences.

In conclusion, the CALE model is a feasible option for modelling choice data to generate the valuation outcome with a large number of attributes. However, it is recommended that more attention should be paid to the distributional difference instead of just evaluating the log-likelihood value and significance statistics.

Table 25 CALE model UK and Australian general characteristics

Country		CALE DCE_{TTO}	CALE DCE-Death
UK	Log-likelihood statistics[1]	-9023.27	-18437.17
	Prob > chi2[2]	0.165	0.119
	Pseudo R-squared[3]	0.165	0.119
	Logical inconsistent	0	0
	Insignificant 10%	0	0
	Utility of health state with all items at level 1	1	1
	Utility of health state with all items at level 2	0.810	0.776
	Utility of health state with all items at level 3	0.527	0.481
	Utility of health state with all items at level 4	-0.005	-0.073
	Utility of health state with all items at level 5	-0.636	-0.787
	Scale length[4]	1.636	1.787
	Mid-point to length[5]	11.6%	12.5%
AUS	Log-likelihood statistics[1]	-9631.47	-19105.93
	Prob > chi2[2]	0.165	0.119
	Pseudo R-squared[3]	0.135	0.106
	Logical inconsistent	0	0
	Insignificant 10%	0	0
	Utility of health state with all items at level 1	1	1
	Utility of health state with all items at level 2	0.666	0.687
	Utility of health state with all items at level 3	0.581	0.476
	Utility of health state with all items at level 4	0.029	-0.107
	Utility of health state with all items at level 5	-0.439	-0.685
	Scale length[4]	1.439	1.685
	Mid-point to length[5]	23.2%	18.6%

Notes: [1]The Log likelihood statistics is the value of final model. This value cannot be directly compared between models as they used different survey data.
[2]The model significance is evaluated through chi-square statistics. The model is significant if the value is less than 0.05
[3]Pseudo R-squared summarizes the proportion of variance explained by the independent variables. A larger R-squared statistic indicates better explanatory power. This value cannot be directly compared between models as they used different survey data and model function.
[4]Scale length is the difference between utility values for states with level 1 for all dimensions and level 5 for all dimensions
[5]Midpoint to length is the value assessed by dividing the difference between utility value for states with level 1 for all dimensions and level 3 for all dimensions by the scale length.

7.5 Preference heterogeneity

7.5.1 MNL regression with correlation terms

Models examining preference heterogeneity across sociodemographic factors of gender, age, carer and cared status, as well as the general health condition, with the DCE_{TTO} standard order design are reported in Table 1 and 2, Appendix F. With both the UK and Australian samples, gender, carer status, cared status and general health have an insignificant main effect, while age, care status and cared status interaction terms are significant. With the UK dataset, older age, respondents cared by others and respondents with better health had lower disutility with the attribute levels, while older age, respondents who did not take care of others or be cared by other people had lower disutility.

7.5.2 Latent class analysis

CAIC and BIC are minimized with 4 classes (with 100 iterations due to computing capacity) with UK sample (Table 3a, Appendix F). The last class, with a strong ordered preference with all level attributes, includes 45.5% of all respondents (Table 4, Appendix F). CAIC and BIC are minimized with 6 classes (with 100 iterations due to computing capacity) with Australian sample (Table 3b, Appendix F). The sixth class explains the decision pattern of 34.3% of all respondents, including more older, male, non-carer/cared healthier respondents (Table 4, Appendix F).

We assume a within-class constant time preference in the analysis, similar to the time preference with the conditional logit regression. Among the six classes with UK data set, Class IV weighed more for the independent duration effect compared with Class I and Class II but had a similar time preference value compared with Class III. With the Australian data set, the Class VI weighed more for the independent duration effect compared with Class II and Class V, but had a lower time preference value compared with Class I, Class III and Class IV. Australian respondents had a more varied time preference compared with the latent class analysis results with the UK data.

7.6 Discussion

This chapter presented the data analysis results to systematically test the feasibility of valuing longer HWB measure with 13 attributes, using the DCE_{TTO} and DCE-Death design. DCE generated high-quality online responses that only 13/19 (UK/Australia) responses were excluded due to response quality. Compared with other DCE health state valuation studies[318], the proportion of respondents reporting hard to understand was relatively small in this study (3% for EQ-HWB DCE valuation V.S. 12.3% DMD-QoL DCE valuation), and the proportion of respondents reporting that they found it was hard to make a choice was slightly higher (45% vs 40%). By examining the model performance, the research was able to conclude that it was feasible to generate value set on QALY scale with the longer HWB measure using the DCE method.

To my knowledge, this study is the first DCE health state valuation study comparing stated preferences using ranking and regressed preferences using DCE. Although there was a large amount of information presented within each DCE task, participants' stated preference and regressed preference revealed high consistency. The stated preferences using the rank task were more aligned with the modelled DCE_{TTO} results than the modelled DCE-Death results. However, we should always be cautious about the stated ranking responses since the ranking task was a posterior test that occurred in the survey after the DCE questions. A different result might be possible if the question design or position was altered in the survey.

By comparing the Model 1 and Model 2 results in each country, the analysis outcome suggested that there were some ordering effects on the utility decrements for some attribute levels of Seeing, Hearing, Sleep, Lonely, Depression and Control. The utility value distributions had more similarities than differences across the two samples where either health or wellbeing attributes appeared first. One explanation for the finding, that was contradicted to what has been found previously, was that the 13 attributes were selected from a single HWB measure, whereas other studies used clusters from different measures[202]. Each of the EQ-HWB attributes included was

unique and non-overlapping dimension, after systematic attribute generation process (by the E-QALY project) and attribute selection process (See Chapter 4). Respondents primarily considered the most important pieces of information instead of the order, as suggested by DCE valuation study using the QLU-C10D measure[275]. Putting a cluster of health or wellbeing attributes first did not lead to an overall significant preference difference.

There was a significant design effect when comparing the DCE_{TTO} and DCE-Death results from the perspective of attribute preference and the utility distribution, where respondents in the DCE-Death group were more cautious about ranking the health state as WTD but had lower utility value once the utility was classified as WTD. The Wald test outcome indicated that the majority of the anchored coefficients were different. Some key aspects may be the reasons for the value set difference, and these refer to: (1) whether it was implicit or explicit that the respondent is stating that a state is WTD; and (2) modelling strategies. The DCE_{TTO} respondents implicitly considered death, whereas the DCE-Death tasks asked respondents to compare each health state directly to being dead. Respondents may be reluctant to place themselves as explicitly preferring death (namely a hesitation to death)[321]. The modelling strategies, especially with varied utility function or data set, influenced the scale length and preference of attributes[322]. Respondents to the DCE-Death task gave larger utility decrements to physical/mental health attributes than wellbeing attributes.

One potential concern with the DCE-Death design was the lack of WTD responses for some respondents, as some respondents do not select any state as WTD. The study revealed that 10% to 20% of respondents never selected any state as WTD, which was lower than the proportion reported in EQ-5D TTO valuation study[119]. With not all participants valuing a state WTD, the health state utilities for the WTD might not reflect the general preference of non-selectors[279]. A similar concern with the DCE_{TTO} design was respondents used heuristics when completing DCE_{TTO} tasks. The self-reported decision-making question asked respondents about their main decision-making strategy, indicating that around 35% respondents considered all of the

information and around 20% of respondents considered the duration attribute only when making their choices. By using a constant duration level, the proportion of respondents mainly considering the life-extending value decreased to around 4%. However, it is hard to make a judgement on whether anchoring the WTD health state with smaller proportion of respondents, as with DCE-Death sample, or generating the value set with larger proportion of potentially heuristic answers, as with the DCE_{TTO} sample, is better[271].

By comparing the data from UK and Australia, we hypothesis that countries were similar in terms of aggregated values and other country-specific circumstance. Some characteristics of preference (i.e., the HWB relative weight) difference were identified, but there are striking similarities in decision and distribution of utility values. Little variance on the health state utility level was explained by the country-level difference. The conclusion was consistent with comparison study with the EQ-5D-5L value set[279].

Apart from the additive model with each level coefficients coded as dummy variables, the CALE model is another feasible option for valuing the HWB measure. An advantage with the CALE model is that there are no insignificant and logically inconsistent factors. Researchers do not need to use the consistent utility models to “absorb’ the insignificant levels or attributes. The CALE models share some similar characteristics in utility distribution as the additive utility models, which is left-skewed with longer tail, and the decision characteristic that deficiency is considered more serious with WTD states (disutility increases incrementally for more severe states). However, the additive model produced a level decrement “kink”, whereas the CALE model flattened the distribution to being more linear across the severity levels and averaged the disutility across the attributes[323, 324]. The CALE model also required strong assumptions about the single impact for different types of attributes that were used in EQ-HWB, which were potentially unrealistic.

Models were estimated to examine preference heterogeneity through the inclusion of interaction terms with key social-demographic characteristics and also through the use of a latent class model. There was evidence of preference heterogeneity across all dimensions to different degrees. Preference for the dimensions of both health and wellbeing differed in different groups of respondents, which emphased the importance of sample representativeness for generating a social HWB preference value set in the future.

Chapter 8 Discussion and Conclusion

8.1 Main findings

This thesis reported an international study of using the DCE method in valuing the EQ-HWB measure, with the research aim to assess the feasibility of generating a value set anchored on 0 to 1 death-full health scale. The study provided qualitative and quantitative evidence on the feasibility of DCE, focusing on the influence of attribute order, design and task formatting, and sample country on the HWB preference. The study was conducted in four main stages.

In stage I (Chapter 1-3), a scoping review and literature review were conducted to summarize the implications of current health state valuation methods and DCE design strategies from both theoretical and practical perspectives. Over 90% of DCE_{TTO} studies explicitly assumed linear time preference[123]. Other methodology options, including a corrected time preference in DCE_{TTO} data regression, ranking design and personal/adaptive preference elicitation, were elicited. For the choice set selection, efficient design was the preferred approach due to its ability to select the most efficient choice sets using informative priors. However, efficient designs typically relied on priors from published value sets or small-sample pilot studies to select the most efficient choice sets. The EuroQol group identified 4 patterns of preference from different regions with the EQ-5D value sets, indicating using published prior information from different country may lead to incorrect predictions and inefficient designs[279].

The second stage (Chapter 4) involved the systematic selection of EQ-HWB attributes to construct a DCE survey with lower cognitive burden, reduced collinearity between attributes, and fewer dependent error terms. The aim was to test the feasibility of valuing more than 9 attributes but fewer than 25 attributes to avoid generating low-quality data. By reviewing all of the EQ-HWB attributes with 7 criteria, 13 attributes selected for the further valuation design. The attribute selection followed DCE design

principles[191, 197] and illustrated a novel perspective of valuing a “EQ-HWB-S bolt-on” in the future: starting with the EQ-HWB-S measure and picking more important EQ-HWB attributes to generate a bolt-on value set, as the health economists do with the EQ-5D bolt-on studies[325].

The third stage (Chapter 5) validated the DCE designs using qualitative data from four semi-structured focus group discussions. I developed a triplet DCE design approach to the anchoring latent value with relative position to death, as an alternative to the DCE_{TTO} task. Participants interpreted the DCE and attribute information as expected. As reported by other qualitative studies[108], participants often used heuristics to make decisions, such as focusing on several important attributes or only considering the duration levels. Several studies have considered the rationale of anchoring with dead using ranking method[4, 326] or duration levels with quantitative survey data before[124, 129], but this study was the first to discuss and head-to-head compare the implications of varied design in a structured way. For good to mild health states the paired design with duration and a triplet design with death may not make difference, but when comparing WTD states or states close to dead, duration attribute played a critical role in locating the health states on a QALY scale more than HWB informaiton.

Finally, Chapters 6 and 7 documented the DCE survey development and application, deriving a value set of 13 HWB attributes from a large sample of 4022 respondents in the UK and Australia. The data demonstrated the feasibility of using DCE_{TTO} and DCE-Death to generate stable and comparable population value sets. Results indicated that DCE design and population influenced HWB distribution and relative preferences, while the order of information had a weak effect on the level coefficient. However, a large number (around 20 per 53 terms in each model of latent coefficients) were insignificances, logical inconsistencies or a combination of both, especially attributes with the frequency levels. The statistical insignificance may partly be due to sample size or the level wording, as there was some evidence that the frequency levels contributed more than severity levels. The qualitative findings (Chapter 5) indicated

that there were some difficulties distinguishing between levels 2 “only occasionally” and levels 3 “sometimes” which may have had an impact on the DCE results.

The relative preference of health and wellbeing attributes were from 4.04 to 7.10 in this study, of which sleep, exhausted and concentration valued less. The EQ-HWB pilot valuation in the UK found similar relative preference that wellbeing attributes were less valued compared to physical health attributes as well[167]. However, the preference with physical health contradicted with UK WTP preference research using ICECAP-A, where the weight of capability and wellbeing attributes (measure) were 0.7-1.6 times larger than the value of health[327]. A PROMIS-29 valuation study, evaluating the preference with emotional wellbeing and general health, found respondents valued physical function, anxiety and sleep more[136]. Different relative weights may because of dimensions/attributes valued and the DCE method used, as well as understanding and implicit classification of attributes. The normative frameworks for wording the descriptive terms and DCE information presentation, as proposed by Hausman and Baker *et al.*, needed a broader academic and public engagement before finalization[328, 329].

A methodology contribution of this study is the use of CALE model with DCE data and the application of the DCE-Death design as an alternative to the DCE_{TTO} task. The CALE mode assumes a constant level effect across all attributes[309], particular attention paying to the significance of worst level and the existence of cross-attribute effect. Throughout the 4 regressions with CALE model described in Section 7.4, all of the level 5 attributes are significant, and the cross-attribute levels are similar with different samples. This empirical evidence supports a deeper exploration of the reliability and comparability of constrained attribute effect assumption with a more flexible constrained attribute function.

The DCE-Death design, a concise ranking valuation method with no varied time, evaluates disutility relative to death, providing reliable anchoring without time trade-offs[120]. It is intuitive to use and easy to choose, arguably making the anchoring to

death more straightforward. Another advantage is the proportion of respondents considered all information increased from 32%(DCE_{TTO}) to 42% (DCE-Death), given that the triplet design respondents had no chance to focus on the length of life change. Although the DCE-Death design has the concern that triplet design considered questions with varied decision strategy[126], DCE-Death should be distinguished from the BWDCE[181, 330], as the third state dead was rarely considered best in most of the cases, nor does the DCE-Death introduced a health third scenario as BWDCE do. Our regressions indicated the best-and-worst data in a single DCE-Death question provided similar attribute ranking as the DCE_{TTO} design in two countries.

8.2 Recommendations for HWB long measure valuation

There is a plethora of valuation methodology researchers exploring choice methods in constructing value set for long measures[67, 109, 219]. However, a gold standard has not been established[331, 332], or any conclusive guideline for the design and anchoring. The empirical evidence in this study contributes to this work stream and translates the ranking valuation design into a triplet DCE design with death. Overall, both DCE_{TTO} and DCE-Death are versatile and pragmatic tools to elicit health state values with HWB measure. The DCE-Death has demonstrated advantages in data quality, level monotonicity, wellbeing attribute significance and consistency to stated preference. However, the DCE-Death has generated larger utility range and more right-skewed utility distribution. As most of the published studies uses the DCE_{TTO} method, a feasibility evaluation study using DCE_{TTO} method increases the comparability with other empirical evidence. By controlling the design effect, the proportion of significant but non-monotonic attributes, utility range and model performance can be compared. However, researchers should pay more attention to duration level selection by testing the levels with qualitative respondents, instead of directly “borrowing” from classic designs, is necessary and more guaranteed.

So far, it is arbitrary to make a judgement that the DCE-Death is feasible with valuing long HWB measures including the EQ-HWB, but the design has provided an alternative

anchoring option. The anchoring with duration in DCE_{TTO} has been explained as "willingness to sacrifice duration for health"[44]. One question is whether "a willingness to sacrifice" should be equal to "willingness to accept dead state". The WTD non-responses constitutes the answer: 12% - 23% responses refused to consider dead state as WTD explicitly. The WTD perspective has spurned debates that continued today[271]: should we consider the WTD preference explicitly or implicitly from individual respondents. A direct comparison with death can inform the wellbeing effect on terminal health state QALY declines better, increasing the data quality and attribute significance[333]. Besides, the proportion of individual respondents never selecting WTDs provides information on the rationale for using negative utility value in the economics evaluation.

This study explored the valuation study design with a systematic mixed method. By evaluating and selecting the proper HWB attributes, all of the valuated attributes had at least one attribute level significant with all of the regression models. However, one thing that could be considered simultaneously with attribute selection was the level performance. A larger number of frequency levels proved to be insignificant or non-monotonic, which influenced the overall model performance. However, this result should not be interpreted independently: most of the wellbeing attributes were described by frequency levels. Respondents reported understanding issue/inconsistency with the wellbeing attributes and tend to be considered less important than health aspects that posed significant influence on daily life. The target of this research is to provide insights on the preference influence of design factors, and any psychometric property discussion, including the wording of levels, is beyond the scope of this research. Frequency levels in EQ-HWB should not be regarded as "more problematic" compared with the severity levels with the given evidence. On the contrary, the comprehensive psychometric evidence provided by E-QALY project and the following EQ-HWB studies more thoroughly supported the rationality of using frequency levels.

From the methodology perspective, qualitative consultations in this study had implications for understanding the role key design factors play in the process of forming DCE preferences and design follow-up questions. It is recommended for future valuation studies to consider attribute and level performances in the DCE attribute selection, and collect qualitative evidence to support survey design.

8.3 Recommendations for future research

This study plans to do some further analysis with the given data sets. The first exploration is analysing the DCE_{TTO} choice data with optimized time preference method. Marcel Yonker's *et al* proposed the net value of time[334] and a correction of time preference DCE with non-linear Bayesian method[126], as two options to consider the common non-linear time preference with general public. However, one inconsistency for the time preference correction is the design and data modelling assumption, whereas using a non-linear factor in the analysis is contradicted with the efficiency calculation[236]. Besides, the non-linear time preference challenges the conceptual foundation of DCE_{TTO} that error term is linear in episodic RUM. By collecting the self-reporting decision making strategy and time preference choices in the follow-up question, the DCE_{TTO} data in this study can be modelled with time preference segmentation characteristics, instead of the demographic characteristics, in the less demanding heterogeneity model[335]. This method can be a more concise modelling strategy to understand and produce time-preference corrected QALY.

Secondly, a comprehensive research program can be devised to understand the influence on HWB preference by using alternative models, including logit model with HWB interaction terms, CALE model with varied level effect and GARBAGE model – a mixed logit regression. Evaluating the utility models is beyond validating the regression insignificant and logically inconsistency with single levels, but to explore the potential explanation power and predictive power. It is predicted that varied utility models can lead to different effects on mild and extreme states[129]. From an econometric point of view, it would be interesting to define the scope of mild state and

consider whether piecewise regression with EQ-HWB preference is applicable, and discuss the prediction accuracy with self-reported health states (more familiar to respondents) and extreme states (less familiar)[279].

Thirdly, further research can be taken to use stated preference and decision-making strategy information in the heterogeneity models. With the DCE-Death data, one assumption held with the baseline analysis is that the respondents employed similar decision strategy for selecting the best health state and the worst health state. Mixed evidence is revealed with this topic: the assumption of consistent decision was likely to be the ideal situation[336] and the best-worst DCE tasks had lower consistency[337], where social care preference evaluation found consistent preference[338]. This should be explored further with the collected UK and Australian DCE data. The decision-making strategy information can be used in the latent class modelling.

A key challenge to use stated preference and ranking of HWB attributes to make model consistency conclusion is whether they ought to be consistent. Although the regression and stated preference provided consistent results with the DCE-Death design and with the top-five ranked attributes with DCE_{TTO}, further qualitative exploration is required to understand the endogenous of stated preference and its relationship with the DCE questions.

8.4 Limitations

The research has a number of limitations. First, evidence-based study design can be improved. This study conducted the item selected based on the initial psychometric evidence generated by the EQ-HWB development stage from 2018 to 2022. However, some of the item wording and regional psychometric evidence has been updated in the following research (e.g., the confirmatory factor analysis has been re-conducted in 2024 and the exploratory factor analysis results are confirmed to be unproblematic{Zhang, 2024 #947}). The item selection may exclude items that should be included with the updated evidence. Besides, although there was no collinearity after item selection, the item selection approach adopted might not be the most

suitable, as a new round of exploratory factor analysis to explore the structure of selected items provides more information for the appropriateness of the used measure. With the 13 items, this study provided evidence on the feasibility of DCE method but the valuation feasibility of whole EQ-HWB measure still need further consideration.

While using DCE in HWB valuation is practically feasible, the small-sample qualitative evidence does not provide evidence with issues with the relative insignificance of wellbeing attributes persist. All of the DCE samples have been recruited from an existing panel, which might not be fully representative of general public experience with DCE and it is hard to ensure the challenges of bots and fraudulent responses. The respondents have completed either DCE_{TTO} or DCE-Death tasks, indicating decision-making strategies and preference with HWB attributes could be different in each group. Addressing these issues would require a comprehensive program of qualitative research and sampling with face-to-face recruitment.

Secondly, this study reported here uses less common choice set selection of a generator design and uses a novel technique of DCE-Death. The results should be interpreted with caution since this is to my knowledge the only study using these combinations. Future research is recommended to compare the DCE_{TTO} and DCE-Death results and whether CALE model can be extended with varied level effects.

Thirdly, due to the large number of health state combinations, it was impossible to calculate the proportion of WTD health states by considering all of the possible combinations. The Monte Carlo method was used to estimate proportions of WTD states, and this procedure may have a random sequence generation bias. Due to the process of simulation being very time consuming, undertaking five million simulations was the largest number that could be undertaken to inform this PhD submission.

Finally and most importantly, preference heterogeneity is not fully explored. The reported regression outcomes are based on the conditional logit model, where homogeneity preference among different populations holds. To provide information about preference heterogeneity, this study presented the regression outcomes with

latent class models and mixed logit models that considered the demographic factors of respondents displayed in this thesis. The informational data collected in the follow-up questions, including time preference, relative preference of attributes and attitude towards death, can be used to generate a robust value set following Bayesian discrete choice estimation framework[339] and censored value set with WTD states[271]. The societal preference of wellbeing reflects the aggregated preference but not allows unique preference of some individual groups[340]. This information is “wasted” but will be analysed and discussed in the future study with the same dataset. Our method and dataset provided a chance to quantify the influence of individual factors.

8.5 Conclusion

This study conducted a comprehensive mixed method study to test the feasibility of valuing health and wellbeing with varied DCE designs, including DCE_{TTO} with putting health information first, DCE_{TTO} with wellbeing information first and DCE-Death design. The data analysis found that all of the three designs generated value set on QALY scale, with insignificant HWB attributes identified by additive utility function. DCE design and sampling country had significant influence on attribute relative preference, but insignificant influence on the general trend of utility values. DCE-Death design had a smaller number of non-monotonicity and insignificance than DCE_{TTO} design, regardless of information order. CALE model generated a value set with all single level coefficients significant, affecting the mild state utility value distribution.

This study does not mark the end of research on DCE HWB valuation methodology exploration but lays the foundation for further investigation into this topic. The classic efficient design, DCE_{TTO} method, and modelling functions remain feasible for HWB valuation. New pragmatic methods introduced in this study facilitate the construction of preferences with a possibility of fewer inconsistencies, greater significance, and reduced heuristic. Future research suggestions were made.

Appendix

Appendix A: DCE Design, Literature Review and EQ-HWB Measures

Appendix B: Focus group consultation Topic Guide

Appendix C: DCE Survey Design

Appendix D: Two-step Cross-Attribute Level Effect (CALE)

Appendix E: Data Analysis Result, by design, by country and by utility function

Appendix F: Preference Heterogeneity analysis result, by model

Appendix A: DCE design, literature review and EQ-HWB measures

Table A-1: different forms of optimal designs

	Mathematical definition	Preference interpretation	Disadvantage
D-efficient design	maximize the determinant of the information matrix |X'X|	minimize the generalized variance of the parameter with given prior	need accurate prior values
C-efficient design	minimizes the variance of best linear unbiased estimator of $c^T\beta$ with linear OLS regression	minimize the marginal effect variance of utility	constrained c-optimal design always used instead of classical c-optimality
A-efficient design	minimize the trace of the inverse of the information matrix over a specified set of design points (a given model).	minimize the average variance of the parameter estimates based on a pre-specified model	lack of sample
G-efficient design	minimize the maximum prediction variance $d=x'(X'X)^{-1}x$ over a specified set of design points (a given model).	select best likelihood estimation with the given determined points and variance	lack of sample
V/I-efficient design	minimize the average prediction variance $d=\frac{Var(x')}{N_X}$ over a specified set of design points.	select best likelihood estimation with the given determined points and variance	lack of sample

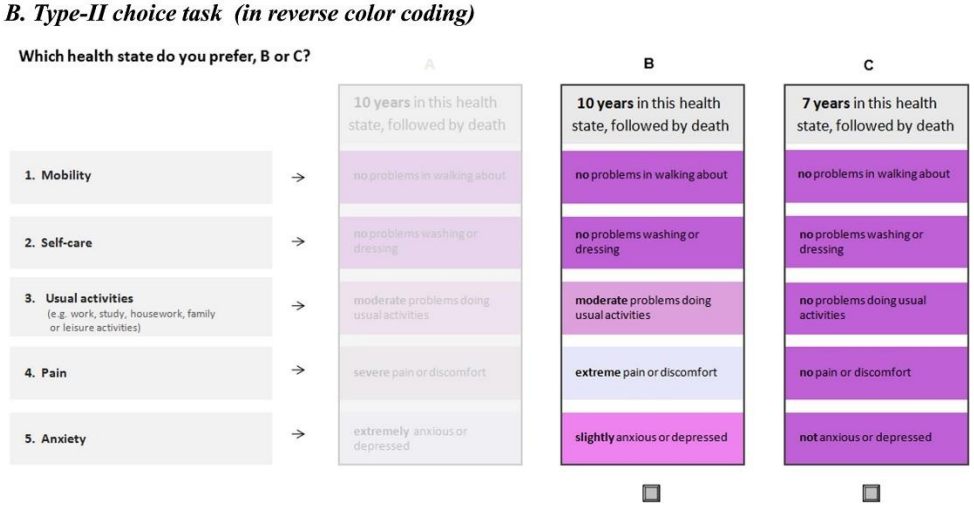
Figure A-1: example of feasible DCE design

a. DCE with duration format

	Health scenario A	Health scenario B
	Slight problems in walking about Moderate problems washing or dressing yourself Severe problems doing your usual activities Severe pain or discomfort Severely anxious or depressed	Slight problems in walking about Severe problems washing or dressing yourself Unable to do your usual activities Extreme pain or discomfort Not anxious or depressed
	Live for 5 years and then die	Live for 10 years and then die
Which scenario do you think is better?	☐	☐

Source: Bansback, N., Hole, A. R., Mulhern, B., & Tsuchiya, A. (2014). Testing a discrete choice experiment including duration to value health states for large descriptive systems: addressing design and sampling issues. *Social science & medicine*, *114*, 38-48.

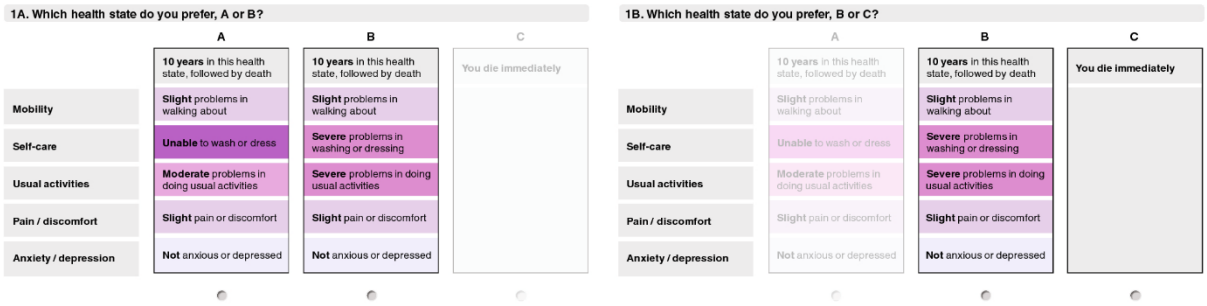
b. the two-way triplet design with perfect health and imperfect health in 10 years



Source: Jonker, M. F., Attema, A. E., Donkers, B., Stolk, E. A., & Versteegh, M. M. (2017). Are

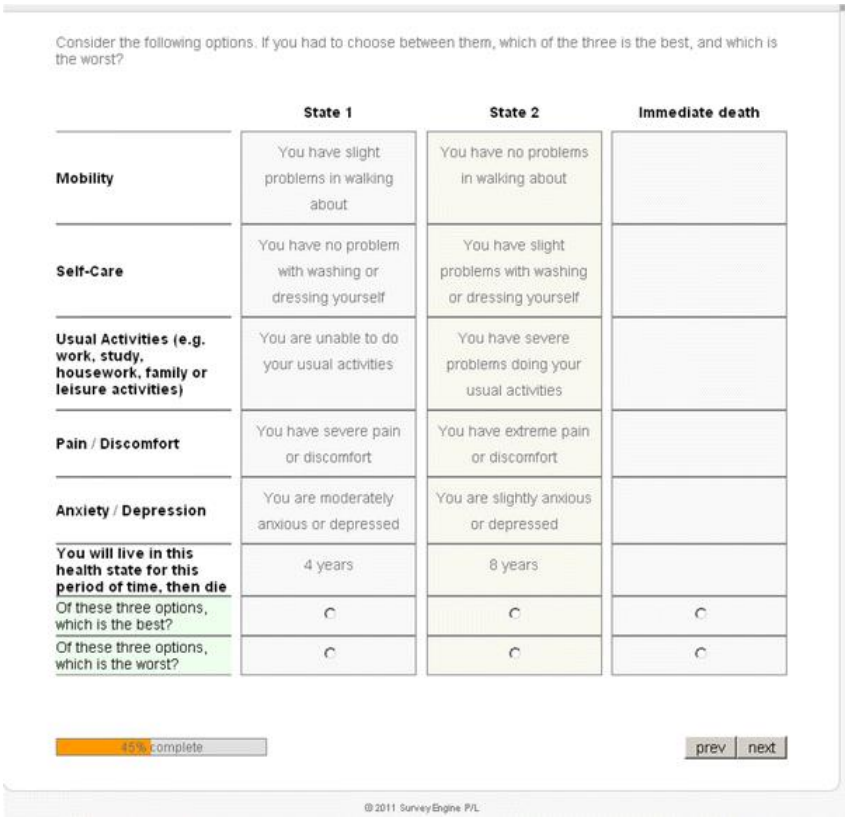
health state valuations from the general public biased? A test of health state reference dependency using self-assessed health and an efficient discrete choice experiment. *Health economics*, *26*(12), 1534-1547.

c. two-way triplet design with immediate death



Source: Lim, S., Jonker, M. F., Oppe, M., Donkers, B., & Stolk, E. (2018). Severity-stratified discrete choice experiment designs for health state evaluations. *Pharmacoeconomics*, *36*, 1377-1389.

d: triplet design with immediate death and varied duration



Source: Norman, R., Cronin, P., & Viney, R. (2013). A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Applied health economics and health policy*, *11*, 287-298.

Table A-2: The EQ-HWB measure (2022 UK version)

*In the **last 7 days**:*		**No difficulty**	**Slight difficulty**	**Some difficulty**	**A lot of difficulty**	**Unable**
1.	How much difficulty did you have seeing? *(using e.g. glasses or contact lenses if you normally use them)*					
2.	How much difficulty did you have hearing? *(using e.g. hearing aids if you normally use them)*					
3.	How much difficulty did you have getting around inside and outside? *(using e.g. a walking stick or wheelchair if you normally use them)*					
4.	How much difficulty did you have doing day-to-day activities? *(e.g. working, shopping, housework)*					
5.	How much difficulty did you have washing, using the toilet, getting dressed, eating, or caring for your appearance?					

*In the **last 7 days**, did you:*		**None of the time**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
6.	have problems with your sleep?					

7.	feel exhausted?					
8.	feel lonely?					
9.	feel that people did not support you?					

*In the **last 7 days**, did you:*		**None of the time**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
10.	have trouble remembering?					
11.	have trouble concentrating or thinking clearly?					
12.	feel anxious?					
13.	feel unsafe? *(e.g. fear of falling, physical harm, abuse)*					
14.	feel frustrated?					
15.	feel sad or depressed?					
16.	feel you had nothing to look forward to?					
17.	feel you had no control over your day-to-day life? *(e.g. had no choice to do things or have*					

	things done for you as you liked and when you wanted)					
18.	feel unable to cope with day-to-day life?					

*In the **last 7 days**:*		**None of the time**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
19.	Did you feel accepted by others? *(e.g. felt like you were able to be yourself and that you belonged)*					
20.	Did you feel good about yourself?					
21.	Could you do the things you wanted to do?					

*In the **last 7 days**, did you:*	**None of the time**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
22. have physical pain?					

23.	Please select **one** response to describe how much physical pain you had in the **last 7 days**. Did you have:	
	no physical pain?	
	mild physical pain?	
	moderate physical pain?	
	severe physical pain?	
	very severe physical pain?	

*In the **last 7 days**, did you:*	**None of the time**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
24. have physical discomfort? *(e.g. feeling sick, breathless, itching) (not including pain)*					

25.	Please select **one** response to describe how much physical discomfort you had in the **last 7 days**. Did you have:	
	no physical discomfort?	
	mild physical discomfort?	
	moderate physical discomfort?	
	severe physical discomfort?	
	very severe physical discomfort?	

Table A-3: The EQ-HWB measure (2022 UK version)

*In the **last 7 days**:*	**No difficulty**	**Slight difficulty**	**Some difficulty**	**A lot of difficulty**	**Unable**
1. How much difficulty did you have getting around inside and outside? *(using e.g. a walking stick or wheelchair if you normally use them)*					
2. How much difficulty did you have doing day-to-day activities? *(e.g. working, shopping, housework)*					

*In the **last 7 days**, did you:*	**None of the time**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
3. feel exhausted?					
4. feel lonely?					

*In the **last 7 days**, did you:*	**None of the time**	**Only occasionally**	**Sometimes**	**Often**	**Most or all of the time**
5. have trouble concentrating or thinking clearly?					
6. feel anxious?					
7. feel sad or depressed?					

8.	feel you had no control over your day-to-day life? *(e.g. had no choice to do things or have things done for you as you liked and when you wanted)*					

9.	Please select **one** response to describe how much physical pain you had in the **last 7 days**. Did you have:	
	no physical pain?	
	mild physical pain?	
	moderate physical pain?	
	severe physical pain?	
	very severe physical pain?	

Table A-4: The EQ-HWB measure (2022 UK version)

Method Key Words:	discrete choice experiment, discrete choice experiments, DCE, conjoint analysis
	离散选择实验， DCE， 联合分析 (translated)
Measurement-related Key Words:	Preference based measure, PBM, EQ-5D, euroqol, SF-6D, Multiattribute utility instrument, MAUI, Utility measure, health related quality of life, quality of life, preferences, health state valuation, valuation, choice experiments, choice modelling
	效用量表，基于偏好的效用量表， EQ-5D，欧洲五维健康量表，euroqol，SF-6D， 六维健康调查简表，多维效用量表，MAUI，健康相关生命质量， 健康相关生活质量， 生活质量， 生命质量，偏好，健康偏好， HRQOL， 健康效用， 健康效用积分体系， 健康状态评价， 效用测量 (translated)

Table A-5: Study categorization

Study	Year	Categorization		Characteristics		Measure
		Data source	Research objective	Country[1]		
Al Shabasy, et al.[341]	2022	Primary	Value set development	Egypt	General public	EQ-5D-5L
Andrade, et al.[342]	2020	Primary	Value set development	French	General public	EQ-5D-5L
Augustovski et al. [201]	2020	Primary	Methodology research	Peru	General public	EQ-5D-5L
Bahrampour et al. [191]	2021	Primary	Value set development	Australia	General public	CP-6D
Baji et al. [189]	2020	Primary	Methodology research	Hungary, Poland, Slovenia	General public	CarerQol-7D
Bouckaert et al.[343]	2021	Primary	Value set development	Belgium	General public	EQ-5D-5L
Chemli, et al. [344]	2021	Primary	Value set development	Tunisia	General public	EQ- 5D-3L
Chen, et al. [345]	2021	Primary	Value set development	Australia	General public	QCE
Comans et al. [135]	2020	Primary	Preference comparison	Australia	General public	AD-5D
Dams et al. [188]	2021	Primary	Value set development	Germany	General public	ICECAP-SCM
Doherty et al. [228]	2021	Secondary	Methodology research	Ireland	General public	EQ-5D-5L
Dufresne et al. [252]	2021	Primary	Value set development	Canada	0-17 children and patients	SF-6Dv2
Fenwick et al. [184]	2020	Primary	Value set development	Australia	Patients	DRU-I
Ferreira et al. [346]	2019	Primary	Value set development	Portugal	General public	EQ-5D-5L
Finch et al. [347]	2021	Primary	Value set development	Italy	General public	EQ-5D-5L
Finch et al. [348]	2021	Primary	Value set development	Spain	General public	QLU-C10D
Gamper et al. [349]	2020	Primary	Value set development	Austra, Italy, Poland	General public	QLU-C10D
Gutierrez-Delgado et	2021	Primary	Value set development	Mexico	General public	EQ-5D-5L

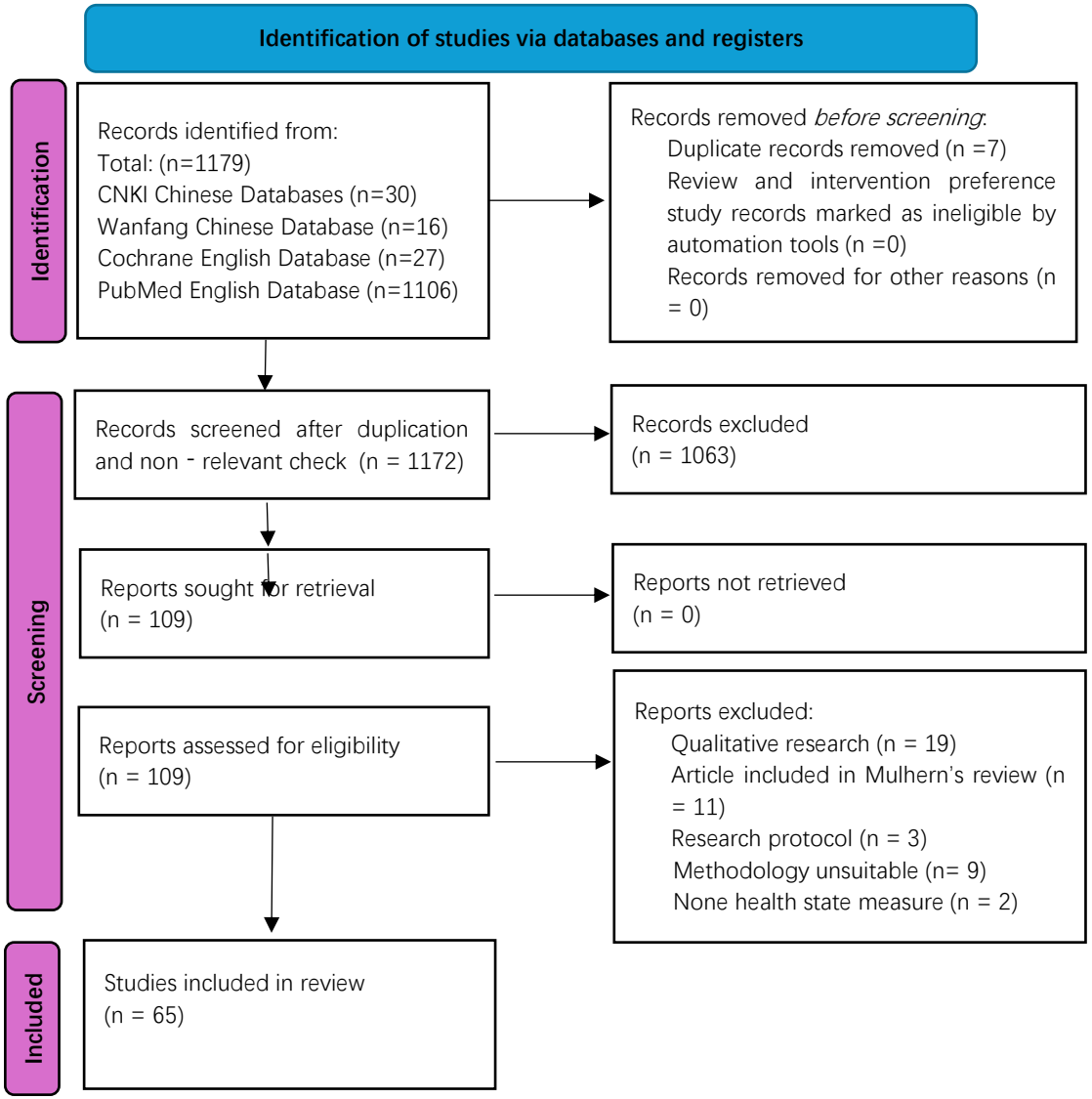
Study	Year	Categorization		Characteristics		Measure
		Data source	Research objective	Country[1]		
al. [350]						
Hansen et al. [351]	2022	Secondary	Methodology research	Norway, Netherlands and United States	General public	EQ-5D-5L
Himmler et al. [125]	2022	Primary	Value set development	Netherland	Elderly people	WOOP
Hoogendoorn et al. [118]	2019	Primary	Preference comparison	Dutch	General public	EQ-5D-5L with bolt-on
Jansen et al. [352]	2021	Primary	Value set development	Dutch	General public	QLU-C10D
Jensen, et al. [353]	2021	Primary	Value set development	Denmark	General public	EQ- 5D-5L
Jiang et al.[354]	2022	Primary	Value set development	US	General public	Neck Disability Index
Jonker, et al. [221]	2019	Primary	Methodology research	Dutch	General public	EQ- 5D-5L
Jyani et al. [139]	2022	Primary	Value set development	India	General public	EQ- 5D-5L
Kemmler et al. [355]	2019	Primary	Value set development	Germany	General public	QLU-C10D (Germany 1/2 versions)
King et al. [356]	2021	Primary	Value set development	Australia	General public	FACT-8D
Krabbe et al. [183]	2020	Primary	Value set development	Hong kong, UK, USA	General public and primary caregivers	IQI
Kreimeier et al. [357]	2022	Primary	Value set development	Germany	General public	EQ-5D-Y
Lim, et al. [126]	2018	Primary	Methodology research	Dutch	General public	EQ- 5D-5L
Ludwig, et al. [358]	2018	Primary	Value set development	Germany	General public	EQ-5D-5L
Malik et al. [225]	2022	Primary	Value set development	Pakistan	General public	EQ-5D-3L
Marten et al. [280]	2020	Secondary	Methodology research	UK	General public	EQ-5D-5L
McTaggart-Cowan et al. [359]	2019	Primary	Value set development	Canada	General public	QLU-C10D
Miguel et al. [360]	2022	Primary	Value set development	Philippines	General public	EQ-5D-5L

Study	Year	Categorization		Characteristics		Measure
		Data source	Research objective	Country[1]		
Mott et al. [182]	2021	Primary	Value set development	UK	General public and adolescence (age 11 to 17)	EQ-5D-Y-3L
Mulhern et al. [202]	2019	Primary	Value set development	Australia	General public	EQ-5D-5L and ASCOT
Mulhern et al. [204]	2020	Primary	Methodology research and Value set development	UK	General public	SF-6Dv2
Nerich et al. [361]	2021	Primary	Value set development	France	General public	QLU-C10D
Norman, et al. [194]	2019	Primary	Value set development	UK	General public	QLU-C10D
O'Hara et al. [186]	2021	Primary	Methodology research	US	General public and people with haemophilia	EQ-5D-5L
Omelyanovskiy et al.[362]	2021	Primary	Value set development	Russia	General public	EQ- 5D-3L
Pattanaphesaj et al. [363]	2018	Primary	Value set development	Thailand	General public	EQ-5D-5L
Pahuta et al. [364]	2021	Primary	Value set development	US	General public	SOSGOQ-8D
Pickard, et al. [365]	2019	Primary	Value set development	US	General public	EQ-5D-5L
Prevolnik and Ogorevc. [366]	2021	Primary	Value set development	Slovenia	General public	EQ- 5D-Y
Ramos-Goñi et al.[367]	2022	Primary	Methodology research and value set development	Spain	General public	EQ- 5D-Y
Ramos-Goñi et al. [368]	2022	Primary	Methodology research	US and UK	General public	EQ- 5D-Y
Ratclife et al. [185]	2022	Primary	Value set development	Australia	Home care and residential care aged people	QOL-ACC

Study	Year	Categorization		Characteristics		Measure
		Data source	Research objective	Country[1]		
Rencz et al. [369]	2022	Primary	Value set development	Hungary	General public	EQ- 5D-Y
Revicki et al. [370]	2021	Primary	Value set development	US	General public	QLU-C10D
Rogers et al. [371]	2022	Primary	Value set development	UK	Adolescents and General public	CARIES-QC-U
Rowen, et al. [197]	2018	Primary	Value set development	UK	General public	HASMID
Rowen, et al. [318]	2021	Primary	Methodology research	UK	General public	DMD-QoL
Roudijk et al. [372]	2022	Primary	Value set development	Netherlands	General public	EQ- 5D-Y
Shafie, et al. [198]	2019	Primary	Value set development	Malaysia	General public	EQ-5D-5L
Shah et al. [195]	2020	Primary	Methodology research	UK	General public	EQ-5D-Y-3L and EQ-5D-3L
Shiroiwa et al. [207]	2021	Primary	Value set development	Japan	General public	EQ- 5D-Y
Sullivan et al. [219]	2020	Primary	Value set development and Methodology research	New Zealand	General public	EQ-5D-5L
Tsuchiya, et al. [373]	2019	Primary	Methodology research	UK	General public	EQ-5D-3L EQ-4D-3L
Voormolen et al. [190]	2020	Primary	Value set development	Italy, Netherland, UK	General public	QOLIBRI-OS
Webb et al. [374]	2020	Primary	Methodology research	UK	General public	EQ-5D-3L
Welie, et al. [375]	2020	Primary	Value set development	Ethiopia	General public	EQ- 5D-5L
Wu et al. [177]	2021	Primary	Value set development	China	General public	SF-6Dv2
Wu et al. [177, 376]	2020	Primary	Methodology research	China	General public	SF-6Dv2
Zhu et al. [377]	2022	Primary	Value set development	China	General public	CQ-11D

Note: 1. Studies included population in more than one country

Figure A-2: Our review was conducted in line with Preferred Reporting Items for Systematic Reviews and Meta-Analyses



Appendix B: Focus group consultation Topic Guide

Valuing Wellbeing alongside Health with the DCE Method

topic guide

23/11/2022

1. Consent, make sure every participant obtained informed consent and oral agreement on site.

2. Introduction (*1.5 Minutes*)
    - Welcome and thank you for joining our group discussion on exploring the study design method.
    - [Introduction to the PhD project and the measure]
    - [Instruction on the focus group objective]

3. Ground Rules (*0.5 Minutes*)

4. Introduction (*5 Minutes*)

5. [Topic one: DCE presentation strategy] (*20 Minutes*)

[*present a dominant example*]. The left column shows the aspect of health or wellbeing which are pretty similar with the information in the questionnaire you have. The middle and right columns show the descriptions, called Life A and Life B. These are levels of the corresponding factor. For example, if there is a "no difficulty" for "In seeing", that means there is no problem with seeing. Aspects that are different for Life A and Life B are highlighted in yellow ("we can see there are six yellow lines) with the remaining 7 all having the same level of severity.

Looking at Life A and Life B, which description would you prefer?
Now it is your time and please tell me your answer once you make up your mind.

6. ICE-BREAKING (*5 Minutes*)

To start with, I'd like to let you get familiar with these aspects by evaluating your own health and wellbeing based on these aspects. Under each heading, please tick the one box that best describes your health or wellbeing in the last 7 days. If there are any questions about the aspect, please raise your hand and I will further explain to you.

Thank you for filling the questionnaire. I could see all of you have pretty good health.

The format is the same as the the warm up question but the two LIFE information is different [*present one DCE choice set with efficient design as example*] As before, there is Life A and B with different description level and differences are highlighted with light yellow colour.

a. If you were asked to select the hypothetical state you prefer, would you be able to make up your decision? How did you make the decision? *[Open discussion, why is that? Ask what do you think of the presentation of the information? Whether or not the fixed attributes provides useful information to assist your decision?]*

b. Does the attribute highlighting (yellow colour) influence your decision-making process?

[Open discussion. If yes, how does it influence? Provide other color-coding examples (i) shades of colours for different levels (ii) "traffic light" colour for different levels, do you prefer light yellow or the other two? Why? Do you think the used colour influence your decision process (color-coding option)?]

[When answer is "NO" move onto]

c. What about other design factors. Do you think the order of descriptive factors influences your decision?

[Open discussion. Does it influence a lot and cause any difficulty in interpreting the information? Do you use any techniques to overcome the influence?]

d. Are there any parts of the description that you find not clear? Which one and why?

[Optional. Summarize points raised before]

e. Which is better / which do you think is better *[Highlighting question]*

Subject important?

7. [Topic two: DCE design strategy] (*24 Minutes, 10 for the first and second comparison each, 4 for question explanation in total*)
Here we presents you with two new designs of comparison questions, which is similar with the first design. Each of the new questions consists of Life A and Life B containing the same statement information, but using a different presentation strategy.
The strategy 2 has two columns, with a full sentence to describe each factor. We can see that factor levels are underlined and the attributes that are different are highlighted with grey.

Presentation strategy 3 has three columns but the first column uses one word to summarize the factor, as a label. And the following two columns are levels. No highlighting colour is used. Asking double negative *[Highlighting question]* also refer the *never feel unsupported* in warm-up.

   a. Which presentation strategy do you prefer, do you think any of the design characteristics in the two new designs helps you to understand the information easier?
      [Open discussion – probe, why is that? Ask for how do they think about having label for each attribute
      What is participants' views on long attribute description]

8. Each of the new comparison consists of Life A and Life B containing the same attribute information as well. Design strategy 4 has a label for each factor. The following two columns provided level information, and we could see factors with the same levels are grouped. For example, we can see in the first part, lonely, sleeping, concentrating and daily activity are all Slight difficulty/only occasionally in Life A and Life B. In the third and fourth part, hearing is No difficulty, and there is Moderate physical pain for both life. The second part contains level information of those factors with different levels. For example, there is no vision problem for life

A and a lot of vision difficulty for life B. Light and deep colour in both lifes represent different levels, the deeper the colour, the more serious the health problem will be.

For presentation strategy 5, we divide a single question into three sub-questions and each sub-question only present part of the information. It consists of only 3~6 factors. The overall question number will be tripled. In another word, fewer information for each question, but the question number is proportionally increased (2~3 times). The first sub-question contains four physical function factors, the second sub-question contains six mental or wellbeing factors, the last sub-question contains sleep, exhausting and pain feeling. The factors are described by a sentence reporting the health problem and its severity, just as the previous design shows. Levels are written with bold text and all together the three questions present the same information as the previous designs.

a. Which presentation strategy do you prefer, or which design strategy helps you to understand the information easier?
[Open discussion – probe, why is that? Ask do you think containing less information in each task, but have more tasks, is better or worse than one task contains more information, but have less task? Why?
How did you make your choice if less information presented? Will you consider attributes that not appear? If yes, how?]

b. Compared with a more complete design, do you think the small aspects enough for making judgement?

9. [Topic three: DCE anchoring] (*26 Minutes, 10 for the first and second comparison each, 6 for question explanation in total*)
Let's look at strategy 6. Life A and Life B presents you with the same information as above. However, there is a duration information for you to consider: imagine you will live with the health problems for the given number of years, then followed by death. The given time varies across the two states. For example, if you select Life A, it means living with the health problems described in Life A for 10

years is better than living with the health problems described in Life B for 4 years. When you are doing the comparison, you need to consider both quality of life and duration of life information. ***[need time control, shorter]***

a. Compared with the comparison in presentation strategy I, where no time information included, do you think the added time information influenced your decision? What do you think of the state with the time in it? *[Ask participants how they consider time with health: do you think the statement and question is clear enough? How do you think about the question in general and how do you make your choice?]*

b. You like the duration on top or bottom? Is it influence your decision process [Highlighting question]

10. Life A consists of more severe mobility problems, mental health issue and pain feelings. Sometimes people may consider a health state is so bad that is worse than death. For example, people may regard persistent vegetative state as worse than death. There are three questions: compare Life A with being dead (from antepenultimate line), compare Life B with being dead and compare Life A and Life B. You need to make the decision and provide your answer on the last three lines. *[Highlighting question]*

c. How do you think about the question in general/Do you think you can always make your choice? Do you think this question is hard to answer? Do you think there is a health state worse than death? *[If answer yes, ask to describe. If answer No, why?]*

d. Which task, presentation strategy 6 and 7, do you prefer to answer?

11. [Topic four: DCE question wording and discrete choice question sets] (*8 Minutes*)

Imagine that you need to complete 12 comparison questions similar to the first task. But this time, there is no one like me to help you to understand.

a. Do you think you can understand what you need to do with my introduction in the first question? Do you think you can understand the question without my explanation at first [If no, ask: *What other background information do you need for understanding the question*

b. There are several examples to ask the task question on the last page. Which expression do you think the clearest and easiest to understand what you should do?
   - *Please consider and imagine living with the two health descriptions below. Then tell us which description you would prefer to live in.*
   - *Please indicate which state is better*

c. Is there anything important that you would like to mention about the task design and understanding?

[Ending Topic]

That is the end of my questions, anything else you would like to bring up?

12. Summary & thank participants, thank the interviewee for their time and useful input to the design of our study. Reaffirm that what we discussed today will remain anonymous.

The presented DCE designs:

	Life A	**Life B**
In seeing	No difficulty	No difficulty
In hearing	No difficulty	No difficulty
In getting around inside and outside	No difficulty	No difficulty
In doing day-to-day activities	Slight difficulty	No difficulty
You feel you have no control over your day-to-day life	Never	Never
You have trouble concentrating/thinking clearly	Only occasionally	Never
You feel anxious	Never	Never
You feel sad/depressed	Never	Never
You feel lonely	Never	Never
You feel unsupported by people	Only occasionally	Never
You have problems with your sleep	Only occasionally	Never
You feel exhausted	Only occasionally	Never
You have physical pain	Mild physical pain	No physical pain
Which would you choose?	◌	◌

	Life A	Life B
	You have **no difficulty** seeing	You have **some difficulty** seeing
	You have **no difficulty** hearing	You have **no difficulty** hearing
	You have **some difficulty** getting around inside and outside	You have **slight difficulty** getting around inside and outside
	You have **slight difficulty** doing day-to-day activities	You have **slight difficulty** doing day-to-day activities
	You **never** feel you have no control over your day-to-day life	You **sometimes** feel you have no control over your day-to-day life
	You **only occasionally** have trouble concentrating/thinking clearly	You **only occasionally** have trouble concentrating/thinking clearly
	You **sometimes** feel anxious	You **often** feel anxious
	You **often** feel sad/depressed	You **only occasionally** feel sad/depressed
	You **only occasionally** feel lonely	You **only occasionally** feel lonely
	You **only occasionally** feel unsupported by people	You **sometimes** feel unsupported by people
	You **only occasionally** have problem with your sleep	You **only occasionally** have problem with your sleep
	You **only occasionally** feel exhausted	You **never** feel exhausted
	You have **moderate** physical pain	You have **moderate** physical pain
Which is better? Life A or B	◌	◌

	Life A	Life B
Vision	No difficulty seeing	Some difficulty seeing
Hearing	No difficulty hearing	No difficulty hearing
Mobility	Some difficulty getting around inside and outside	Slight difficulty getting around inside and outside
Daily activity	Slight difficulty doing day-to-day activities	Slight difficulty doing day-to-day activities
Control	Never feel you have no control over your day-to-day life	Sometimes feel no control over your day-to-day life
Concentrating/ thinking clearly	Only occasionally have trouble concentrating/thinking clearly	Only occasionally have trouble concentrating/thinking clearly
Anxious	Feel anxious sometimes	Often feel anxious
Sad/depressed	Often feel sad/depressed	Feel sad/depressed only occasionally
Loneliness	Feel lonely only occasionally	Feel lonely only occasionally
Support	Feel unsupported by people only occasionally	Feel unsupported by people sometimes
Sleep	Have problem with sleep only occasionally	Have problem with sleep only occasionally
Fatigue	Feel exhausted only occasionally	Never feel exhausted
Pain	Moderate physical pain	Moderate physical pain
Which is better? Life A or B	◌	◌

	Life A	Life B
Lonely **Sleeping** **concentrating/thinking clearly** **Daily activity**	Slight difficulty/ only occasionally	

Mobility	Some difficulty	Slight difficulty
Anxious	Sometimes	Often
Sad/depressed	Often	Only occasionally
Control	Never	Sometimes
Unsupported by people	Only occasionally	Sometimes
Feeling exhausted	Only occasionally	Never
Vision	No difficulty	Some difficulty

Hearing	No difficulty

Pain	Moderate physical pain

Which is better? Life A or B	◌	◌

Life A	Life B	Question 1
No difficulty seeing Have problem with sleep **only occasionally** Feel anxious **sometimes** **Slight difficulty** doing day-to-day activities	**Some difficulty** seeing Have problem with sleep **only occasionally** **Often** feel anxious **Slight difficulty** doing day-to-day activities	
◌	◌	**Which is better? Life A or B**

Life A	Life B	Question 2
Never feel you have no control over your day-to-day life **No difficulty** hearing **Some difficulty** getting around inside and outside **Often** feel sad/depressed Feel lonely **only occasionally** Feel unsupported by people **only occasionally**	**Sometimes** feel no control over your day-to-day life **No difficulty** hearing **Slight difficulty** getting around inside and outside Feel sad/depressed **only occasionally** Feel lonely **only occasionally** Feel unsupported by people **sometimes**	
◌	◌	**Which is better? Life A or B**

Life A	Life B	Question 3
Only occasionally have trouble concentrating/thinking clearly Feel exhausted **only occasionally** **Moderate** physical pain	**Only occasionally** have trouble concentrating/thinking clearly **Never** feel exhausted **Moderate** physical pain	
◌	◌	**Which is better? Life A or B**

	Life A	Life B
In seeing	Unable	A lot of difficulty
In hearing	No difficulty	No difficulty
In getting around inside and outside	A lot of difficulty	Some difficulty
In doing day-to-day activities	A lot of difficulty	A lot of difficulty
You feel you have no control over your day-to-day life	Never	Sometimes
You have trouble concentrating/thinking clearly	Often	Often
You feel anxious	Most or all of the time	Often
You feel sad/depressed	Often	Only occasionally
You feel lonely	Often	Often
You feel unsupported by people	Only occasionally	Sometimes
You have problems with your sleep	Most or all of the time	Most or all of the time
You feel exhausted	Only occasionally	Never
You have physical pain	Severe physical pain	Severe physical pain
Better than being dead	◌	◌
Worse than being dead	◌	◌
Which is better? Life A or B	◌	◌

Appendix C: DCE Survey Design

Table C-1: the DCE design samples

a. The DCE_{TTO} paired comparison format example: EQ-HWB order design I

	Life A		Life B
In seeing	No difficulty		Some difficulty
In hearing	No difficulty		No difficulty
In getting around inside and outside	Some difficulty		Slight difficulty
In doing day-to-day activities	Slight difficulty		Slight difficulty
You have problems with your sleep	Only occasionally		Only occasionally
You feel exhausted	Only occasionally		None of the time
You feel lonely	Only occasionally		Only occasionally
You feel unsupported by people	Only occasionally		Sometimes
You have trouble concentrating/thinking clearly	Only occasionally		Only occasionally
You feel anxious	Sometimes		Often
You feel sad/depressed	Often		Only occasionally
You feel you have no control over your day-to-day life	None of the time		Sometimes
You have physical pain	Moderate physical pain		Moderate physical pain
You will live in the state for	**4 years and then die**		**5 years and then die**
Which would you choose? Life A or Life B			

b. The DCE_{TTO} paired comparison format example: wellbeing first order design II

	Life A		Life B
You feel you have no control over your day-to-day life	None of the time		Sometimes
You feel lonely	Only occasionally		Only occasionally
You feel unsupported by people	Only occasionally		Sometimes
You have trouble concentrating/thinking clearly	Only occasionally		Only occasionally
You feel anxious	Sometimes		Often
You feel sad/depressed	Often		Only occasionally
You have problems with your sleep	Only occasionally		Only occasionally
You feel exhausted	Only occasionally		None of the time
In seeing	No difficulty		Some difficulty
In hearing	No difficulty		No difficulty
In getting around inside and outside	Some difficulty		Slight difficulty
In doing day-to-day activities	Slight difficulty		Slight difficulty
You have physical pain	Moderate physical pain		Moderate physical pain
You will live in the state for	**4 years and then die**		**5 years and then die**
Which would you choose? Life A or Life B			

c. The DCE-Death triplet comparison format example: design III

	Life A	Life B	Immediate death
In seeing	No difficulty	Some difficulty	
In hearing	No difficulty	No difficulty	
In getting around inside and outside	Some difficulty	Slight difficulty	
In doing your day-to-day activities	Slight difficulty	Slight difficulty	
You have problems with your sleep	Only occasionally	Only occasionally	
You feel exhausted	Only occasionally	None of the time	
You feel lonely	Only occasionally	Only occasionally	
You feel unsupported by people	Only occasionally	Sometimes	
You have trouble concentrating/thinking clearly	Only occasionally	Only occasionally	
You feel anxious	Sometimes	Often	
You feel sad/depressed	Often	Only occasionally	
You feel you have no control over your day-to-day life	None of the time	Sometimes	
You have physical pain	Moderate physical pain	Moderate physical pain	
You will live in the health state for	**10 years and then die**	**10 years and then die**	
Of these three options, which is the best?			
Of these three options, which is the worst?			

Valuing Health and Wellbeing

Consent Form (example)

This survey will ask respondents to answer comparison questions including death as an option

Please initial the appropriate boxes	**Yes**	**No**
Taking Part in the Project		
I have read and understood the project information above.	☐	☐
I agree to take part in the project. I understand that taking part in the project will include completing the research survey questionnaire which involves comparison questions on health, wellbeing, life expectancy and death.	☐	☐
I understand that choosing to participate as a volunteer in this research, does not create a legally binding agreement, nor is it intended to create an employment relationship with the University of Sheffield and the consulting company SurveyEngine.	☐	☐
I understand that my taking part is voluntary and that I can close the survey at any point, and if I do not complete the survey, my data will not be used. If I complete the survey, my data cannot be withdrawn as it is anonymous.	☐	☐
How my information will be used during and after the project		
I understand that no personal identifying information will be shared with the researcher.	☐	☐
I understand and agree that my answers may be quoted in publications, reports, web pages, and other research outputs in and out of the University of Sheffield. I understand that I will not be named in these outputs.	☐	☐
I understand and agree that other authorised researchers may have access to the anonymous data.	☐	☐
I understand and agree that other authorised researchers may use my anonymous data in publications, reports, web pages, and other research outputs.	☐	☐
So that the information you provide can be used legally by the researchers		
I agree to assign the copyright I hold in any materials generated as part of this project to The University of Sheffield.	☐	☐

This survey will ask respondents to answer comparison questions including death as an option.

Participants are expected to complete the consent form online by clicking the 'agree and understand' button before the online survey start.

Project contact details for further information:

SurveyEngine contact information to be added

Haode Wang: hwang165@sheffield.ac.uk, 30 Regent St, Sheffield City Centre, Sheffield S1 4DA Donna Rowen: d.rowen@sheffield.ac.uk, 30 Regent St, Sheffield City Centre, Sheffield S1 4DA

Participant Information Sheet (example)

Research project title

Valuing health and wellbeing.

1. Invitation to participate

You are invited to take part in a research project. Before you decide whether to participate, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Please contact us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part. Thank you for reading this.

2. What is the project's purpose?

The research project is testing the feasibility of valuing health and wellbeing using a paired comparison method. In each of the questions, we will present you two imaginary states to find out your health and wellbeing preference. This international research will be launched in both the UK and Australia. The findings will be used to instruct future studies and to inform healthcare resource allocation decisions in both countries.

3. Why have I been chosen?

You have been chosen because you are a member of online panels accessed by consulting company Surveyengine and meet our selection criteria.

4. Do I have to take part?

It is up to you to decide whether or not to take part. If you decide to take part, you will be asked to consent before answering the questions. You can also email the survey organiser SurveyEngine **(**email to be added) to ask for more information about the project. You can leave the survey at any time without any reason, and if you do not finish the survey, your data will not be used.

5. What will happen to me if I take part? What do I have to do?

We would like to invite you to take part in a 15-20 minute online survey. It is expected that you complete the questions by yourself, and you do not need discussion with others. There are no right or wrong answers.

Prior to starting the survey, you will complete a consent form. Please take some time to read through the information sheet and consent form.

In the survey, you will be asked questions about you, such as gender, age, education level and health status. Then there will be 12 questions (including a warm-up question) which each show two descriptions of health and wellbeing with life expectancy, and ask you to choose which you think is better. There will be a practice question that explains these questions.

After the comparison questions, you will be asked what you thought of the survey.

6. What are the possible disadvantages and risks of taking part?

There is no anticipated risk in taking part in this survey, but the paired comparison question will ask you to consider aspects of health, wellbeing, and life expectancy in choosing the answer to the question. If you find some of the questions upsetting, or if you wish to seek advice or reassurance about your own health, then either contact your GP or helpline 1800 022 222. It is okay not to answer a question, or to stop the survey by simply clicking the close button. After withdrawing, the data will not be analysed.

7. What are the possible benefits of taking part?

You will be reimbursed by [panel company SurveyEngine use].

8. Will my taking part in this project be kept confidential?

All the information we collect about you during the survey will be kept strictly confidential and will only be accessible to members of the research team and SurveyEngine.

Anonymised data will be used to support research dissemination, including presentations, reports and/or publications. You will not be able to be identified in any reports or publications.

9. What is the legal basis for processing my personal data?

According to data protection legislation, we are required to inform you that the legal basis we are applying in order to process your personal data is that 'processing is necessary for the performance of a task carried out in the public interest' (Article 6(1)(e)). Further information can be found in the University's Privacy Notice https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

10. What will happen to the data collected, and the results of the research project?

All data collected as part of this study will be held securely and in confidence at the University of Sheffield, and available to the research group and related institutes only. No identifying details (names, contact details) will be collected in the survey.

Anonymised data may be shared and archived online to support a research publication and to enable the fair reuse of the data by interested researchers.

We intend to publish the results of this research project in an academic journal and present it at academic conferences. If you would like a copy of the results, please email the survey organizer SurveyEngine **(email to be added)**. You will not be identifiable in these outputs and the results will not be automatically shared with you if you do not actively contact the research team.

11. Who is organising and funding the research?

This research is funded by the EuroQol Research Foundation.

12. Who is the Data Controller?

The University of Sheffield will act as the Data Controller for this study. This means that the University is responsible for looking after your information and using it properly.

13. Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by the Sheffield Centre for Health and Related Researchand the Curtin Human Research Ethics Committee, Australia.

14. What if something goes wrong and I wish to complain about the research or report a concern or incident?

You can withdraw or leave the survey at any time if you feel uncomfortable. You can report any discomfort and distress to Haode Wang or the survey organizer.

If you are dissatisfied with any aspect of the research and wish to make a complaint, please contact the supervisor (Professor Donna Rowen) in the first instance. If you feel your complaint has not been handled in a satisfactory way you can contact the Head of the Division of Population Health (Dr Louise Preston, l.r.preston@sheffield.ac.uk). If the complaint relates to how your personal data has been handled, you can find information about how to raise a complaint in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

If you wish to make a report of a concern or incident relating to potential exploitation, abuse or harm resulting from your involvement in this project, please contact the project's Designated Safeguarding Contact (Professor Donna Rowen, d.rowen@sheffield.ac.uk). If the concern or incident

relates to the Designated Safeguarding Contact, or if you feel a report you have made to this Contact has not been handled in a satisfactory way, please contact the Head of the Division of Population Health (Dr Louise Preston, l.r.preston@sheffield.ac.uk) and/or the University's Research Ethics & Integrity Manager (Lindsay Unwin; l.v.unwin@sheffield.ac.uk).

15. Contact for further information

Apart from survey organizer SurveyEngine **(email to be added),** if you would like to find out more about this research project, you can contact:

Mr Haode Wang
Sheffield Centre for Health and Related Research, University of Sheffield
Regent Court, 30 Regent Street, Sheffield, S1 4DA
hwang165@sheffield.ac.uk

Professor Donna Rowen
Sheffield Centre for Health and Related Research, University of Sheffield
Regent Court, 30 Regent Street, Sheffield, S1 4DA
d.rowen@sheffield.ac.uk

DCE Survey Template (example)

Instructions

[Participant information sheet and consent form link to be added by SurveyEngine after ethics approval]

Please read the consent form carefully before clicking the consent options

The survey has 3 parts.

1) In Part 1, we will ask you 26 questions about you and your health and wellbeing.

2) In Part 2, we will ask you one practice question plus 13 choice questions.

3) In Part 3, we will ask you some follow-up questions about your decision making process and what you thought of the survey.

To fully participate and complete the survey, you need to complete all three parts.

PART 1: Questions about you

The following questions will ask about you and your health and wellbeing. You should answer all of the questions (select at least one option, including the *Prefer not to say* option) to complete this session.

1. Are you:

Male ☐
Female ☐
Other ☐
Prefer not to say ☐

2. What is your age (in years):

[ ]

3. What is your current marital status:

Single ¨
Married/ De facto ¨
Separated/ Divorced ¨
Widowed ¨
Prefer not to say o

4. Which of the following best describes your main activity? Select one box below.
Full-time employed or self-employed ☐
Part-time employed or self-employed..... ☐
Retired ..... ☐
Student ..... ☐
Unemployed ..... ☐
Long-term sickness ..... ☐ _____
Look after family/home..... ☐
Other (please specify) ..... ☐ _____
Prefer not to say ☐

5. What is the highest level of education you have completed?
For Australian version:
Year 11 and below ¨
Year 12 ¨
Certificate (any level including trade certificate) ¨
Diploma/ advanced diploma ¨
Bachelors or honours degree ¨
Post-graduate degree (Masters or Doctorate) ¨
Prefer not to say o

6. What is your annual household/individual income before tax (including benefits)?
For Australian version:
Negative or Zero Income ¨
$1 - $20,799 per year ($1 - $399 per week) ¨
$20,800 - $41,599 per year ($400 - $799 per week) ¨
$41,600 - $77,999 per year ($800 - $1499 per week) ¨
$78,000 - $103,999 per year ($1500 - $1999 per week) ¨
$104,000 - $155,999 per year ($2000- $2999 per week) ¨
$156,000 - $207,999 per year ($3000 - $3999 per week) ¨
$208,000 - $259,999 per year ($4000 - $4999 per week) ¨
$260,000 - $311,999 per year ($5000 - $5999 per week) ¨
$312,000 or more per year ($6000 or more per week) ¨
Prefer not to say or don't know ¨

7. What is your ethnic group?
For Australian version:
Aboriginal and/or Torres Strait Islander origin o
North-West European o
Southern and Eastern European ¨
North-East and South-East Asian ¨
Southern and Central Asian o

North African and Middle Eastern o
Sub-Saharan African ..
Peoples of the Americas ..
Prefer not to say o

8. Are you a parent or guardian for a child or children aged under 18 years?
Yes ☐
No ☐
Prefer not to say ☐

9. Are you a carer for an adult(s) family member or friend (not as a paid job) because of their health or age?
Yes ☐
No ☐
Prefer not to say ☐

10. Are you cared for by other adults (including paid carers) because of your health or your age?
Yes ☐
No ☐
Prefer not to say ☐

11. Are your day-to-day activities limited because of a health problem or disability which has lasted, or is expected to last, at least 12 months?
Include problems related to old age.
Yes, limited a lot ☐
Yes, limited a little ☐
No ☐
Prefer not to say ☐

12. In general, would you say your health is:

Excellent ☐
Very good ☐
Good ☐
Fair ☐
Poor ☐

13. What is your age (in years): [Attention check question. **Drop participant** if different answer to question 2]

14-26. The following questions will ask about your health and wellbeing in general.
These questions are about the last 7 days.

Please answer all questions. There are no wrong or right answers.

Please select **one response** for each question.

PART 2: Choosing which life you prefer to live in

In the following questions, you will be presented with two descriptions of health and wellbeing that last for a certain number of years, called Life A and Life B. We want you to tell us which you would prefer to live in: Life A or Life B.

The aspects of health and wellbeing that may be affected are to do with **your physical health, mental health and wellbeing of daily life**, and are:

- Difficulty seeing (including using, for example, glasses or contact lenses if they are needed)
- Difficulty hearing (including using hearing aids if you usually wear them)
- Difficulty getting around inside and outside (including using any aids you usually use e.g., walking stick, frame or wheelchair)
- Difficulty doing day-to-day activities (e.g., working, shopping, housework)
- Problems with sleeping
- Feeling exhausted
- Feeling lonely
- Feeling unsupported
- Trouble concentrating/thinking clearly
- Feeling anxious
- Feeling sad/depressed
- Feeling you have no control over day-to-day life (e.g., having the choice to do things or have things done for you as you like and when you want)
- Feeling physical pain

When answering the questions, please imagine that you will experience each life for the number of years shown. After that, you will die. Please imagine that death will be quick and pain free and you will not have any other health and wellbeing problems apart from what is mentioned. There will be no additional treatment to make the situation better. Please select either Life A or Life B to show which you prefer.

We will start with a practice question.

Each of the text boxes on the below will appear one at a time and will be formatted clearly in the online survey.

Participant will receive feedback after answering this question.
If participant chose Life A:

You chose **Life A**, which means that you prefer to live for 10 years with:

- **Having some difficulty in getting around inside and outside**
- **Having slight difficulty in doing day-to-day activities**
- **Only occasionally having sleeping problems**
- **Only occasionally feeling exhausted**
- **Only occasionally feeling lonely**
- **Only occasionally feeling unsupported by people**
- **Only occasionally having trouble concentrating/thinking clearly**
- **Sometimes feeling anxious**
- **Often feeling sad/depressed, and**
- **Having moderate level physical pain**

Than to live for 4 years with:

- **Having a lot of difficulty in seeing**
- **Having slight difficulty in getting around inside and outside**
- **Having slight difficulty in doing day-to-day activities**
- **Only occasionally having sleeping problems**
- **Only occasionally feeling lonely**
- **Sometimes feeling unsupported by people**
- **Only occasionally having trouble concentrating/thinking clearly**
- **Often feeling anxious**
- **Only occasionally feeling sad/depressed**
- **Sometimes feeling having no control over the day-to-day life, and**
- **Having moderate level physical pain**

Do you still prefer Life A? (Yes/No)

If yes, proceed to next question
If no, start practice question again

If participant chose Life B:

You chose **Life B**, which means that you prefer to live for **4 years** with:

- **Having a lot of difficulty in seeing**
- **Having slight difficulty in getting around inside and outside**
- **Having slight difficulty in doing day-to-day activities**
- **Only occasionally having sleeping problems**
- **Only occasionally feeling lonely**

- **Sometimes feeling unsupported by people**
- **Only occasionally having trouble concentrating/thinking clearly**
- **Often feeling anxious**
- **Only occasionally feeling sad/depressed**
- **Sometimes feeling having no control over the day-to-day life, and**
- **Having moderate level physical pain**

Than to live for **10 years** with:

- **Having some difficulty in getting around inside and outside**
- **Having slight difficulty in doing day-to-day activities**
- **Only occasionally having sleeping problems**
- **Only occasionally feeling exhausted**
- **Only occasionally feeling lonely**
- **Only occasionally feeling unsupported by people**
- **Only occasionally having trouble concentrating/thinking clearly**
- **Sometimes feeling anxious**
- **Often feeling sad/depressed, and**
- **Having moderate level physical pain**

Do you still want to choose Life B? (Yes/No)

If yes, proceed to the next question
If no, start practice questions again

Survey questions

Now we would like you to answer 13 questions. You cannot skip questions or go back to change your answers (you can quit the survey at any time without penalty). Some descriptions are more difficult to imagine than others, please take your time and consider each option carefully. Your choices will be based on your views, there are no right or wrong answers.
Please keep going until you finish!

[THESE 13 QUESTIONS WILL ALL TAKE THE SAME FORMAT AS THE QUESTION ABOVE, 12 WILL BE FORMAL QUESTIONS AND 1 WILL BE COGNITIVE QUESTION (WITH DOMINATED PAIRS)]

PART 3: After Survey questions:

Decision question after the survey:

In answering the comparison questions above, would you say you (choose one option that best describe how you made your choices):

Considered all of the health and wellbeing aspects all the time ☐

Only considered health and/or wellbeing aspect(s) that I believe to be important ☐

Considered different health and wellbeing aspects each time ☐

Mainly considered the length of time in Life A or B ☐

Considered health and/or wellbeing aspects not presented here

(please specify the health and/or wellbeing aspects) ☐

_____

Other decision method (please specify) ☐ _____

Selected the option at random ☐

Do not know ☐

Attribute importance question:

Among all of the health and wellbeing aspects in the comparison question, which do you think is the most important (you may select more than one):

- Difficulty seeing ☐
- Difficulty hearing ☐
- Difficulty getting around inside and outside ☐
- Difficulty doing day-to-day activities ☐
- Problems sleeping ☐
- Feeling exhausted ☐
- Feeling lonely ☐
- Feeling unsupported ☐
- Trouble concentrating/thinking clearly ☐
- Feeling anxious ☐
- Feeling sad/depressed ☐
- Feeling you have no control over day-to-day life ☐
- Feeling physical pain ☐

Time preference question after the DCE survey (in quantitative):

We would like to know how you would act in the following situation:

Imagine that you have some problems with some difficulty in seeing and hearing, some difficulty in doing day-to-day activities, some fatigue, moderate pain, but no other problems with your health and wellbeing. There are two alternative treatment options (Treatment 1 and 2) available. The effects of the alternative treatments vary with regard to when the illness will occur and how long you will be ill for (you cannot be cured completely). For example, with the Treatment 2, you will be ill starting 5 years from now for 48 days and with treatment 1 you will be ill starting 1 years from now for 20 days. Assuming everything else about the treatments is the same (i.e. severity of the treatment, side effects, costs) which treatment would you prefer?

Treatment 1: you will be ill 1 years from now for 20 days ☐

Treatment 2: you will be ill 5 years from now for 48 days ☐

No preference ☐

Time preference question after the DCE survey (qualitative):

In general, how willing are you to give up something that is beneficial for you today in order to benefit more from that in the future? Please indicate your answer on a scale from 1 to 10 (or don't know), where 1 means you are 'completely unwilling to do so' and a 10 means you are 'very willing to do so'.

Attention check question [Participant will not be dropped]

Which of the following best describes your main activity? Select one.

Full-time employed or self-employed ..... ☐
Part-time employed or self-employed ..... ☐
Retired ..... ☐
Student ..... ☐
Unemployed ..... ☐
Long-term sickness ..... ☐_____
Look after family/home..... ☐
Other (please specify) ..... ☐ _____
Prefer not to say ☐

Feedback questions

Please answer the following questions by selecting **one response** for each question to tell us how you think about the survey and your answers.

	Strongly agree	Agree	Neither agree or disagree	Disagree	Strongly disagree
It was **easy to understand** the questions I was asked?	□	□	□	□	□
I found it **easy to tell the difference** between the two lives I was asked to think about?	□	□	□	□	□
I found it **difficult to decide on my answers** to the questions.	□	□	□	□	□
The **number of choice** questions I had to answer was **appropriate (not too many or too few)**.	□	□	□	□	□
The **amount of information** used to describe the health and wellbeing problems was **appropriate (not too much or too little)**.					
	Very confident	Confident	Neither confident nor unconfident (neutral)	Unconfident	Very unconfident
How confident are you in your answers to the questions where you had to make choices?	□	□	□	□	□

Appendix D: two-step Cross-Attribute Level Effect (CALE)

Cross-Attribute Level Effect (CALE) estimation

1.1 The mathematical assumption of CALE model

This material is to illustrate the mathematical relationship bwtween EQ- addtivie utility model (ADD20) and EQ- 8-factor CALE model (MULT8). The utility function employed an additive conditional logit assumption holding IIA distribution with error terms.

The ADD20:

$$\mu_{ij} = \alpha + \beta'_{see}See'_{ij} + \beta'_{hr}hr'_{ij} + \beta'_{ar}ar'_{ij} + \beta'_{dtd}dtd'_{ij} + \beta'_{slp}slp'_{ij} + \beta'_{exh}exh'_{ij} + \beta'_{ll}ll'_{ij} + \beta'_{spt}spt'_{ij} + \beta'_{tk}tk'_{ij} + \beta'_{axs}axs'_{ij} + \beta'_{dpr}dpr'_{ij} + \beta'_{ctr}ctr'_{ij} + \beta'_{pp}pp'_{ij} + \epsilon_n$$

Where for all of the EQ-HWB factors, the attribute levels are dummy coded. The attributes and correspondent coefficient factors are expressed in vectors. For example, the seeing attribute vector fucntion is:

$$\beta'_{see}See'_{ij}=\beta_{see5}See_{level5} + \beta_{see2}See_{level2} + \beta_{see3}See_{level3} + \beta_{see4}See_{level4}$$

The MULT8:

$$\mu_{ij} = \alpha + (\beta_{see5}See^5_{ij} + \beta_{hr5}hr^5_{ij} + \beta_{ar5}ar^5_{ij} + \beta_{dtd5}dtd^5_{ij} + \beta_{slp5}slp^5_{ij} + \beta_{exh5}exh^5_{ij} + \beta_{ll5}ll^5_{ij} + \beta_{spt5}spt^5_{ij} + \beta_{tk5}tk^5_{ij} + \beta_{axs5}axs^5_{ij} + \beta_{dpr5}dpr^5_{ij} + \beta_{ctr5}ctr^5_{ij} + \beta_{pp5}pp^5_{ij}) + (\beta_{see5}See^4_{ij} + \beta_{hr5}hr^4_{ij} + \beta_{ar5}ar^4_{ij} + \beta_{dtd5}dtd^4_{ij} + \beta_{slp5}slp^4_{ij} + \beta_{exh5}exh^4_{ij} + \beta_{ll5}ll^4_{ij} + \beta_{spt5}spt^4_{ij} + \beta_{tk5}tk^4_{ij} + \beta_{axs5}axs^4_{ij} + \beta_{dpr5}dpr^4_{ij} + \beta_{ctr5}ctr^4_{ij} + \beta_{pp5}pp^4_{ij})L_4 + (\beta_{see5}See^3_{ij} + \beta_{hr5}hr^3_{ij} + \beta_{ar5}ar^3_{ij} + \beta_{dtd5}dtd^3_{ij} + \beta_{slp5}slp^3_{ij} + \beta_{exh5}exh^3_{ij} + \beta_{ll5}ll^3_{ij} + \beta_{spt5}spt^3_{ij} + \beta_{tk5}tk^3_{ij} + \beta_{axs5}axs^3_{ij} + \beta_{dpr5}dpr^3_{ij} + \beta_{ctr5}ctr^3_{ij} + \beta_{pp5}pp^3_{ij})L_3 + (\beta_{see5}See^2_{ij} + \beta_{hr5}hr^2_{ij} + \beta_{ar5}ar^2_{ij} + \beta_{dtd5}dtd^2_{ij} + \beta_{slp5}slp^2_{ij} + \beta_{exh5}exh^2_{ij} + \beta_{ll5}ll^2_{ij} + \beta_{spt5}spt^2_{ij} + \beta_{tk5}tk^2_{ij} + \beta_{axs5}axs^2_{ij} + \beta_{dpr5}dpr^2_{ij} + \beta_{ctr5}ctr^2_{ij} + \beta_{pp5}pp^2_{ij})L_2 + \epsilon_n$$

The estimation of **MULT8** constrained variant of **ADD20** into 13 parameters representing the disutility of level 5 on each of the dimensions. The disutility of other three levels are calculated by multiplying parameters for levels 4, 3 and 2 (L_4、L_3、L_2). This constrained function relies on two strong assumptions: **the first assumption** is that the imperfect levels for all attributes, except the level five, are proportional to level five with the same parameter. A preference interpretation is that the relative preference of each attribute is fixed and will not change with varied levels. The participant's relative preferences on levels are independent of the preference on attributes. This assumption is similar with the preference assumption of Online Elicitation of Personal Utility Functions (OPUF) but different from the assumption of classic DCE utility function, where the preference on each of the attribute level are independent of each other. **The second assumption** is that the coefficient error term of all levels of certain attribute have identical distribution form. The

identical distribution is different from the i.i.d assumption with alternative utility function, where we assumed a identical error term distribution among individuals but not on the attributes coefficients. With the identical distribution form assumption, we assume that all of the attribute levels will be significant if the level 5 is significant, compared to level 1 as the baseline, because the variation range will not cross 0 by multiplying positive L_i values. These two assumptions increase the estimation efficiency and level significance, optimize utility distribution and diminish the possibility of inconsistency. However, the strong assumptions do not allow any insignificant middle levels (if level 5 significant) or significant middle levels (if level 5 insignificant), which may violate the real-world preference with the supportive qualitative or psychometric evidence. A evenly distributed utility values can be different from the standard bimodal utility value distribution. Besides, the estimation uses a non-linear conditional logit regression models for the DCE-TTO data, and methods for fitting such models are less accessible than linear conditional logit regression methods on STATA.

Taking the first two assumptions into the utility function, we can derive the **third assumption** that the coding format of level variables changed from dummy coding to continuous coding. In another word, with the **MULT8** utility function, the dummy coded variables are linearly transferred to some relationship of level 5, instead of independent estimation. We take the health state comparison of State A 4444444444444 and State B 5555555555555. For State A, we have the **ADD20** estimation utility function:

$$\mu_{iA} = \beta_{see4}See_{ij}^4 + \beta_{hr4}hr_{ij}^4 + \beta_{ar4}ar_{ij}^4 + \beta_{dtd4}dtd_{ij}^4 + \beta_{slp4}slp_{ij}^4 + \beta_{exh4}exh_{ij}^4 + \beta_{ll4}ll_{ij}^4 + \beta_{spt4}spt_{ij}^4 + \beta_{tk4}tk_{ij}^4 + \beta_{axs4}axs_{ij}^4 + \beta_{dpr4}dpr_{ij}^4 + \beta_{ctr4}ctr_{ij}^4 + \beta_{pp4}pp_{ij}^4 = (\beta_{see5}See_{ij}^4 + \beta_{hr5}hr_{ij}^4 + \beta_{ar5}ar_{ij}^4 + \beta_{atd5}dtd_{ij}^4 + \beta_{slp5}slp_{ij}^4 + \beta_{exh5}exh_{ij}^4 + \beta_{ll5}ll_{ij}^4 + \beta_{spt5}spt_{ij}^4 + \beta_{tk5}tk_{ij}^4 + \beta_{axs5}axs_{ij}^4 + \beta_{dpr5}dpr_{ij}^4 + \beta_{ctr5}ctr_{ij}^4 + \beta_{pp5}pp_{ij}^4)L_4$$

The dummy coded variable equals to 1 if State A is 4444444444444, then we have:

$$\mu_{iA} = (\beta_{see5} \times 1 + \beta_{hr5} \times 1 + \beta_{ar5} \times 1 + \beta_{atd5} \times 1 + \beta_{slp5} \times 1 + \beta_{exh5} \times 1 + \beta_{ll5} \times 1 + \beta_{spt5} \times 1 + \beta_{tk5} \times 1 + \beta_{axs5} \times 1 + \beta_{dpr5} \times 1 + \beta_{ctr5} \times 1 + \beta_{pp5} \times 1)L_4$$

For State B, we use the same strategy to get the function:

$$\mu_{iB} = (\beta_{see5} \times 1 + \beta_{hr5} \times 1 + \beta_{ar5} \times 1 + \beta_{atd5} \times 1 + \beta_{slp5} \times 1 + \beta_{exh5} \times 1 + \beta_{ll5} \times 1 + \beta_{spt5} \times 1 + \beta_{tk5} \times 1 + \beta_{axs5} \times 1 + \beta_{dpr5} \times 1 + \beta_{ctr5} \times 1 + \beta_{pp5} \times 1)$$

Then $\mu_{iA} = \mu_{iB} \times L_4$, we can generalize this function to any health state J such as 3333333444444 by using vector of L_s, to express the utility of Health state J $\mu_{iJ} = \mu_{iB} \times L_n'$ as

a function of State 5555555555555. Each of the X_{ijx}^{n}=$X_{ijx}^{5} \times L_{in}$. As a result, each of the measure dimension should be expressed with one variable level 5 for estimating the β'_{x5}, and the other levels are expressed as $X_{ijx}^{5} \times L_{in}$, which is a linear relationship with the level 5 dummy coded variable (1 or 0). However, as introduced above, this has changed the independent dummy coding relationship between the variables. A new variable for each dimension is needed, where the values for each levels are 0-for level 1 (no disutility), L2-for level 2 (level 5 multipled by L_2), L3-for level 3 (level 5 multipled by L_3), L4-for level 2 (level 5 multipled by L_4), and 1 for level 5.

To transfer the model of TTO data on R and fit conditional logit regression model with DCE data on STATA, we need to change the function to linear function estimation. Note that all of the factors in **MULT8** are dummy coded so that we can neither directly change this function to shorten the function with the assumption that $See_{leveli,i\neq5} = f(See_{level5})$, nor change this to the format to directly estimating 13+4 (13 level-5 coefficients of all of the factors + 4 level coefficients) because of the probability estimation characteristic of conditional logit where a "function of function" is not allowed. In another word, the estimation of $\beta_{see5}See_{ij}^{5} + \beta_{hr5}hr_{ij}^{5} + \beta_{ar5}ar_{ij}^{5} + \beta_{dtd5}dtd_{ij}^{5} + \beta_{slp5}slp_{ij}^{5} + \beta_{exh5}exh_{ij}^{5} + \beta_{ll5}ll_{ij}^{5} + \beta_{spt5}spt_{ij}^{5} + \beta_{tk5}tk_{ij}^{5} + \beta_{axs5}axs_{ij}^{5} + \beta_{dpr5}dpr_{ij}^{5} + \beta_{ctr5}ctr_{ij}^{5} + \beta_{pp5}pp_{ij}^{5}$ cannot be combined with the estimation of $(\beta_{see5}See_{ij}^{4} + \beta_{hr5}hr_{ij}^{4} + \beta_{ar5}ar_{ij}^{4} + \beta_{dtd5}dtd_{ij}^{4} + \beta_{slp5}slp_{ij}^{4} + \beta_{exh5}exh_{ij}^{4} + \beta_{ll5}ll_{ij}^{4} + \beta_{spt5}spt_{ij}^{4} + \beta_{tk5}tk_{ij}^{4} + \beta_{axs5}axs_{ij}^{4} + \beta_{dpr5}dpr_{ij}^{4} + \beta_{ctr5}ctr_{ij}^{4} + \beta_{pp5}pp_{ij}^{4})L_4$ or other sub-optimised levels because they transferred the estimation to $\beta_{see5}See_{ij}^{5} + \beta_{see5}See_{ij}^{4} \times L_4 = \beta_{see5}(See_{ij}^{5} + See_{ij}^{4} \times L_4)$ where the degree of freedom for estimation (1+1=2) exceeds the dummy coded factor degree of freedom in a utility function (only 1) in estimation. In another word, the requirement of **MULT8** estimation requests prior information for L_4、L_3、L_2 to change the function into a format with single factor for each coefficient estimation.

1.2 Consider the situation of paired comparison DCE data

However, if we let The **ADD20 = The MULT8** (which is our final target and the two functions simulates the same utility outcome) with the assumption of **MULT8** that the factor of sub-optimal levels are proportionally similar among the attributes. Then we can get the function (take Level 4 as an example):

$$(\beta_{see4}See_{ij}^{4} + \beta_{hr5}hr_{ij}^{4} + \beta_{ar4}ar_{ij}^{4} + \beta_{dtd4}dtd_{ij}^{4} + \beta_{slp4}slp_{ij}^{4} + \beta_{exh4}exh_{ij}^{4} + \beta_{ll4}ll_{ij}^{4} + \beta_{spt4}spt_{ij}^{4} + \beta_{tk4}tk_{ij}^{4} + \beta_{axs4}axs_{ij}^{4} + \beta_{dpr4}dpr_{ij}^{4} + \beta_{ctr4}ctr_{ij}^{4} + \beta_{pp4}pp_{ij}^{4}) = (\beta_{see5}See_{ij}^{4} + \beta_{hr5}hr_{ij}^{4} + \beta_{ar5}ar_{ij}^{4} + \beta_{dtd5}dtd_{ij}^{4} + \beta_{slp5}slp_{ij}^{4} + \beta_{exh5}exh_{ij}^{4} + \beta_{ll5}ll_{ij}^{4} + \beta_{spt5}spt_{ij}^{4} + \beta_{tk5}tk_{ij}^{4} + \beta_{axs5}axs_{ij}^{4} + \beta_{dpr5}dpr_{ij}^{4} + \beta_{ctr5}ctr_{ij}^{4} + \beta_{pp5}pp_{ij}^{4})L_4$$

Then we get:

$$\frac{(\beta_{see4}See_{ij}^{4}+\beta_{hr5}hr_{ij}^{4}+\beta_{ar4}ar_{ij}^{4}+\beta_{dtd4}dtd_{ij}^{4}+\beta_{slp4}slp_{ij}^{4}+\beta_{exh4}exh_{ij}^{4}+\beta_{ll4}ll_{ij}^{4}+\beta_{spt4}spt_{ij}^{4}+\beta_{tk4}tk_{ij}^{4}+\beta_{axs4}axs_{ij}^{4}+\beta_{dpr4}dpr_{ij}^{4}+\beta_{ctr4}ctr_{ij}^{4}+\beta_{pp4}pp_{ij}^{4})}{(\beta_{see5}See_{ij}^{4}+\beta_{hr5}hr_{ij}^{4}+\beta_{ar5}ar_{ij}^{4}+\beta_{dtd5}dtd_{ij}^{4}+\beta_{slp5}slp_{ij}^{4}+\beta_{exh5}exh_{ij}^{4}+\beta_{ll5}ll_{ij}^{4}+\beta_{spt5}spt_{ij}^{4}+\beta_{tk5}tk_{ij}^{4}+\beta_{axs5}axs_{ij}^{4}+\beta_{dpr5}dpr_{ij}^{4}+\beta_{ctr5}ctr_{ij}^{4}+\beta_{pp5}pp_{ij}^{4})} = L_4$$

Let all of the attribute levels equal to 4, which is the health state utility of 4^{13}, we get

$$\beta_{see4}See_{ij}^{4}+\beta_{hr5}hr_{ij}^{4}+\beta_{ar4}ar_{ij}^{4}+\beta_{dtd4}dtd_{ij}^{4}+\beta_{slp4}slp_{ij}^{4}+\beta_{exh4}exh_{ij}^{4}+\beta_{ll4}ll_{ij}^{4}+\beta_{spt4}spt_{ij}^{4}+\beta_{tk4}tk_{ij}^{4}+\beta_{axs4}axs_{ij}^{4}+\beta_{dpr4}dpr_{ij}^{4}+\beta_{ctr4}ctr_{ij}^{4}+\beta_{pp4}pp_{ij}^{4} = \beta_{see4}+\beta_{hr5}+\beta_{ar4}+\beta_{dtd4}+\beta_{slp4}+\beta_{exh4}+\beta_{ll4}+\beta_{spt4}+\beta_{tk4}+\beta_{axs4}+\beta_{dpr4}+\beta_{ctr4}+\beta_{pp4}$$

Same for the function below, with which we can get that of the same attribute :

$$\beta_{see5}See_{ij}^{4}+\beta_{hr5}hr_{ij}^{4}+\beta_{ar5}ar_{ij}^{4}+\beta_{dtd5}dtd_{ij}^{4}+\beta_{slp5}slp_{ij}^{4}+\beta_{exh5}exh_{ij}^{4}+\beta_{ll5}ll_{ij}^{4}+\beta_{spt5}spt_{ij}^{4}+\beta_{tk5}tk_{ij}^{4}+\beta_{axs5}axs_{ij}^{4}+\beta_{dpr5}dpr_{ij}^{4}+\beta_{ctr5}ctr_{ij}^{4}+\beta_{pp5}pp_{ij}^{4} = \beta_{see5}+\beta_{hr5}+\beta_{ar5}+\beta_{dtd5}+\beta_{slp5}+\beta_{exh5}+\beta_{ll5}+\beta_{spt5}+\beta_{tk5}+\beta_{axs5}+\beta_{dpr5}+\beta_{ctr5}+\beta_{pp5}$$

Using the health state 4444444444444 in the function, where all of the dummy coded X =1, and taking the above two functions into the L_4 function, a prior relation between β_4^{13} and β_5^{13}, which recorded by L_4, equals to $\sum\beta_4'/\sum\beta_5'$. Using the same method (with health state of 3333333333333 and 2222222222222), we can get L_3 and L_2 to get the two parameters equal to $\sum\beta_3'/\sum\beta_5'$ and $\sum\beta_3'/\sum\beta_5'$. All of the β_{ijx}' can be derived by conditional logit regression and the three L parameters can be interpreted as the relative importance of certain level to level 5. The level significance will not influence the relative importance calculation.

If we take the duration interaction into account, we can notice that for the

$$\mu_{ij}=\beta_{duration}Duration_{ij}+[297]$$

Considering the fact that each factor of the **The ADD20** is also interacted with duration and the intercept term (duration) is not related with L_4 or L_3 or L_2, then L_4 will not change with or without duration if the duration is coded as continuous variable. However, if duration is dummy coded, which indicates a non-linear time preference or discounted time value assumption taken, then the $L_4=\frac{\sum\beta_4'}{\sum\beta_5'}\times\frac{\overline{C_{n\,(\mathbf{ADD20})}^{1}}}{C_{n(\boldsymbol{MULT\,8})}^{1}}$ (for the dummy coded time variable), where the n is the levels of duration attribute. If the discounted time method used, the function remains $\sum\beta_4'/\sum\beta_5'$.

By taking the function of $\sum\beta_4'/\sum\beta_5'$ into the **MULT8** function, we can get the utility function is to multiply each of the **Level 4** dummy coded independednt variable with a constant number, which is derived from **The ADD20** regression. Note that this two-stage regression (**ADD20** regression to get the constant number with the assumption that the β_4'、β_3'、β_2' have linear relation with β_5') are a compromised method for one-step estimation, under the conditional logit (binary) model framework. A conditional logit model estimates the

linear function of odds ratio of binary choice data:

$$In\left(\frac{P_{n1}}{1-P_{n2}}\right) = \beta_{duration}(Duration_{i1} - Duration_{i2}) + \sum_{i=number\ of\ attributes}^{1} \beta_{ni}\ (X_{ni1} - X_{ni2})$$

If we take the L4, L3, L2 into account, then the estimation would not be generalized linear estimation, where the interaction of parameters can not be estimated within the conditional logit framework:

$$In\left(\frac{P_{n1}}{1-P_{n2}}\right) = \beta_{duration}(Duration_{i1} - Duration_{i2}) + \sum_{i=number\ of\ attributes}^{1} \beta_{5i}\ (X_{ni1} - X_{ni2}) \times L4/L3/L2$$

As a result, the two-stage estimation is the we that we should use. We estimate the L4, L3, L2 parameters in the first stage and estimate the 13 level-five attribute β_s in the second stage.

1.3 Consider the situation of triplet comparison DCE data

If we use the triplet comparison data as paired comparison data, which means to de-composite the choice data to Health State A- Health State B, Health State A-Death and Health State B-Death, then the above two-step regression can be applied in the same way. For Health State A and Health State B, we have utility functions:

$$\begin{aligned}\mu_{ij} = Dummy_{death} &+ (\beta_{see5}See_{ij}^{5} + \beta_{hr5}hr_{ij}^{5} + \beta_{ar5}ar_{ij}^{5} + \beta_{dtd5}dtd_{ij}^{5} + \beta_{slp5}slp_{ij}^{5} + \beta_{exh5}exh_{ij}^{5} + \beta_{ll5}ll_{ij}^{5} + \beta_{spt5}spt_{ij}^{5} + \beta_{tk5}tk_{ij}^{5} + \beta_{axs5}axs_{ij}^{5} + \beta_{dpr5}dpr_{ij}^{5} + \beta_{ctr5}ctr_{ij}^{5} + \beta_{pp5}pp_{ij}^{5}) \\ &+ \left(\beta_{see5}See_{ij}^{4} + \beta_{hr5}hr_{ij}^{4} + \beta_{ar5}ar_{ij}^{4} + \beta_{dtd5}dtd_{ij}^{4} + \beta_{slp5}slp_{ij}^{4} + \beta_{exh5}exh_{ij}^{4} + \beta_{ll5}ll_{ij}^{4} + \beta_{spt5}spt_{ij}^{4} + \beta_{tk5}tk_{ij}^{4} + \beta_{axs5}axs_{ij}^{4} + \beta_{dpr5}dpr_{ij}^{4} + \beta_{ctr5}ctr_{ij}^{4} + \beta_{pp5}pp_{ij}^{4}\right)L_4 \\ &+ (\beta_{see5}See_{ij}^{3} + \beta_{hr5}hr_{ij}^{3} + \beta_{ar5}ar_{ij}^{3} + \beta_{dtd5}dtd_{ij}^{3} + \beta_{slp5}slp_{ij}^{3} + \beta_{exh5}exh_{ij}^{3} + \beta_{ll5}ll_{ij}^{3} + \beta_{spt5}spt_{ij}^{3} + \beta_{tk5}tk_{ij}^{3} + \beta_{axs5}axs_{ij}^{3} + \beta_{dpr5}dpr_{ij}^{3} + \beta_{ctr5}ctr_{ij}^{3} + \beta_{pp5}pp_{ij}^{3})L_3 \\ &+ (\beta_{see5}See_{ij}^{2} + \beta_{hr5}hr_{ij}^{2} + \beta_{ar5}ar_{ij}^{2} + \beta_{dtd5}dtd_{ij}^{2} + \beta_{slp5}slp_{ij}^{2} + \beta_{exh5}exh_{ij}^{2} + \beta_{ll5}ll_{ij}^{2} + \beta_{spt5}spt_{ij}^{2} + \beta_{tk5}tk_{ij}^{2} + \beta_{axs5}axs_{ij}^{2} + \beta_{dpr5}dpr_{ij}^{2} + \beta_{ctr5}ctr_{ij}^{2} + \beta_{pp5}pp_{ij}^{2})L_2 \\ &+ \epsilon_n\end{aligned}$$

L4, L3, L2 are constant numbers from step 1. For state death, we have:

$$\mu_{ij} = Dummy_{death} + \epsilon_n$$

Then the estimation of β_5 do not change.

If we use the triplet comparison data as triplet comparison data, the estimation will be under ranked-order logit or other multinorminal logit frameworks. The estimation will be to estimate values of parameters that make choice probabilities consistent with observed choices using the maximum likelihood estimation. In this estimation, we will not use a triplet comparison estimation regarding the with-DEATH data as a three-profile DCE comparison, instead we uses the full ranking information from the best to the worst (ranking

1 to 3). The ranking analysis follows the Allison and Christakis (1994) and McFadden (1973) frameworks. This framework is to estimate each ranking number preference with a sequential conditional logit model. To get a ranking order γ_i' with 3 alternatives, we need to have the utility order:

$$U'_{i\gamma_{i1}} > U'_{i\gamma_{i2}} > U'_{i\gamma_{i3}}$$

The probability of achieving a ranking order γ_i' is to estimate the product of 2 (J-1) probabilities:

$$\Pr[\gamma_i';\beta] = \Pr\left[U'_{i\gamma_{i1}} > U'_{i\gamma_{i2}} > U'_{i\gamma_{i3}}\right] = \prod_{j=1}^{2} \frac{\exp\left(V_{irij}\right)}{\sum_{l=j}^{3} \exp\left(V_{iril}\right)} = \frac{e^{V_{i1}}}{e^{V_{i1}} + e^{V_{i2}} + e^{V_{i3}}} \times \frac{e^{V_{i2}}}{e^{V_{i2}} + e^{V_{i3}}}$$

A log likelihood estimation of the above function is:

$$\ln L_{n=1} = (V_{i1} + V_{i2}) - (\ln(e^{V_{i1}} + e^{V_{i2}} + e^{V_{i3}}) + \ln\left(e^{V_{i2}} + e^{V_{i3}}\right))$$

Note that the utility function in paired comparison DCE data will replace V_{i1}, which is then maximuzed with respect to the coefficient vectors. By replacing the **ADD20** utility function with the **MULT8** function, estimation can be conducted within the same framework and relative importance transformations to get Li has the same mathematical properties after $e^{V_{i1}}$ transformation. A two-step estimation or one-step estimation use same likelihood estimation strategy. We choose to use the two-step estimation here to maintain a model consistency and prevent the interaction terms $\beta_s \times L_i$.

Appendix E: Data analysis result, by design, by country and by utility function

Table E-1: UK sample self-report health condition

		ANOVA statistic (Prob>F)[1]	Cohort 1		Cohort 2		Cohort 3		Overall	
			Number	Proportion	Number	Proportion	Number	Proportion	Number	Proportion
			627	100.00%	600	100.00%	776	100.00%	2003	100.00%
Health status										
Day-to-day activities limitation										
	Yes, limited a lot	0.12 (0.88)	50	7.97%	40	6.67%	50	6.44%	140	6.99%
	Yes, limited a little		117	18.66%	127	21.17%	156	20.10%	400	19.97%
	No		455	72.57%	428	71.33%	565	72.81%	1,448	72.29%
	Prefer not to say		5	0.80%	5	0.83%	5	0.64%	15	0.75%
General health										
	Excellent	0.49 (0.61)	59	9.41%	71	11.83%	94	12.11%	224	11.18%
	Very good		208	33.17%	186	31.00%	229	29.51%	623	31.10%
	Good		196	31.26%	194	32.33%	278	35.82%	668	33.35%
	Fair		141	22.49%	124	20.67%	147	18.94%	412	20.57%
	Poor		23	3.67%	25	4.17%	28	3.61%	76	3.79%

EQ-HWB health status self-report										
Vision	No difficulty	1.49 (0.23)	416	66.35%	378	63.00%	499	64.30%	1,293	64.55%
	Slight difficulty		151	24.08%	154	25.67%	188	24.23%	493	24.61%
	Some difficulty		54	8.61%	56	9.33%	67	8.63%	177	8.84%
	A lot of difficulty		6	0.96%	11	1.83%	20	2.58%	37	1.85%
	Unable			0.00%	1	0.17%	2	0.26%	3	0.15%
Hearing	No difficulty	0.44 (0.65)	495	78.95%	482	80.33%	625	80.54%	1,602	79.98%
	Slight difficulty		89	14.19%	80	13.33%	111	14.30%	280	13.98%
	Some difficulty		32	5.10%	30	5.00%	26	3.35%	88	4.39%
	A lot of difficulty		11	1.75%	7	1.17%	13	1.68%	31	1.55%
	Unable			0.00%	1	0.17%	1	0.13%	2	0.10%
Getting around	No difficulty	0.82 (0.44)	509	81.18%	500	83.33%	644	82.99%	1,653	82.53%
	Slight difficulty		60	9.57%	56	9.33%	66	8.51%	182	9.09%
	Some difficulty		34	5.42%	26	4.33%	45	5.80%	105	5.24%
	A lot of difficulty		21	3.35%	16	2.67%	18	2.32%	55	2.75%
	Unable		3	0.48%	2	0.33%	3	0.39%	8	0.40%
Day-to-day activities	No difficulty	0.01 (0.99)	463	73.84%	433	72.17%	562	72.42%	1,458	72.79%
	Slight difficulty		94	14.99%	98	16.33%	124	15.98%	316	15.78%
	Some difficulty		38	6.06%	47	7.83%	57	7.35%	142	7.09%
	A lot of difficulty		26	4.15%	18	3.00%	30	3.87%	74	3.69%
	Unable		6	0.96%	4	0.67%	3	0.39%	13	0.65%
Sleep	None of the time	0.23 (0.79)	173	27.59%	155	25.83%	212	27.32%	540	26.96%
	Only occasionally		206	32.85%	197	32.83%	254	32.73%	657	32.80%
	Sometimes		115	18.34%	122	20.33%	157	20.23%	394	19.67%
	Often		94	14.99%	84	14.00%	97	12.50%	275	13.73%

	Most or all of the time		39	6.22%	42	7.00%	56	7.22%	137	6.84%
Exhausted	None of the time	0.91 (0.40)	194	30.94%	166	27.67%	218	28.09%	578	28.86%
	Only occasionally		188	29.98%	182	30.33%	229	29.51%	599	29.91%
	Sometimes		128	20.41%	130	21.67%	171	22.04%	429	21.42%
	Often		82	13.08%	94	15.67%	112	14.43%	288	14.38%
	Most or all of the time		35	5.58%	28	4.67%	46	5.93%	109	5.44%
Lonely	None of the time	1.21 (0.30)	304	48.48%	302	50.33%	369	47.55%	975	48.68%
	Only occasionally		120	19.14%	122	20.33%	143	18.43%	385	19.22%
	Sometimes		115	18.34%	106	17.67%	161	20.75%	382	19.07%
	Often		60	9.57%	44	7.33%	63	8.12%	167	8.34%
	Most or all of the time		28	4.47%	26	4.33%	40	5.15%	94	4.69%
Support	None of the time	0.46 (0.63)	352	56.14%	323	53.83%	414	53.35%	1,089	54.37%
	Only occasionally		124	19.78%	129	21.50%	173	22.29%	426	21.27%
	Sometimes		89	14.19%	88	14.67%	114	14.69%	291	14.53%
	Often		48	7.66%	33	5.50%	40	5.15%	121	6.04%
	Most or all of the time		14	2.23%	27	4.50%	35	4.51%	76	3.79%
Concentration/thinking clearly	None of the time	0.76 (0.47)	302	48.17%	271	45.17%	347	44.72%	920	45.93%
	Only occasionally		138	22.01%	153	25.50%	190	24.48%	481	24.01%
	Sometimes		111	17.70%	93	15.50%	137	17.65%	341	17.02%
	Often		63	10.05%	59	9.83%	72	9.28%	194	9.69%
	Most or all of the time		13	2.07%	24	4.00%	30	3.87%	67	3.34%
Anxious	None of the time	0.05 (0.95)	246	39.23%	234	39.00%	300	38.66%	780	38.94%
	Only occasionally		150	23.92%	148	24.67%	185	23.84%	483	24.11%
	Sometimes		112	17.86%	108	18.00%	154	19.85%	374	18.67%
	Often		89	14.19%	74	12.33%	84	10.82%	247	12.33%
	Most or all of the time		30	4.78%	36	6.00%	53	6.83%	119	5.94%
Depression	None of the time	1.13	280	44.66%	254	42.33%	301	38.79%	835	41.69%

	Only occasionally	(0.32)	144	22.97%	150	25.00%	213	27.45%	507	25.31%
	Sometimes		101	16.11%	107	17.83%	136	17.53%	344	17.17%
	Often		79	12.60%	62	10.33%	83	10.70%	224	11.18%
	Most or all of the time		23	3.67%	27	4.50%	43	5.54%	93	4.64%
Control	None of the time	0.24 (0.79)	345	55.02%	298	49.67%	399	51.42%	1,042	52.02%
	Only occasionally		120	19.14%	143	23.83%	184	23.71%	447	22.32%
	Sometimes		88	14.04%	98	16.33%	107	13.79%	293	14.63%
	Often		49	7.81%	38	6.33%	63	8.12%	150	7.49%
	Most or all of the time		25	3.99%	23	3.83%	23	2.96%	71	3.54%
Physical Pain	No physical pain	0.92 (0.40)	287	45.77%	254	42.33%	319	41.11%	860	42.94%
	Mild physical pain		227	36.20%	229	38.17%	311	40.08%	767	38.29%
	Moderate physical pain		89	14.19%	87	14.50%	119	15.34%	295	14.73%
	Severe physical pain		18	2.87%	22	3.67%	24	3.09%	64	3.20%
	Very severe physical pain		6	0.96%	8	1.33%	3	0.39%	17	0.85%

[1] Analysis of variance (ANOVA) tested the variance among the three cohorts to test if their baseline health condition was significantly different. The F-statistic, combined with the degree of freedom information, significant with a 5% significance level.

Table E-2: Australian sample self-report health condition

		ANOVA statistic (Prob>F)[1]	Cohort 1		Cohort 2		Cohort3		Overall	
			Number	Proportion	Number	Proportion	Number	Proportion	Number	Proportion
			632	100.00%	604	100.00%	783	100.00%	2019	100.00%
Health status										
Day-to-day activities limitation										
	Yes, limited a lot	0.40 (0.67)	59	9.34%	59	9.77%	77	9.83%	195	9.66%
	Yes, limited a little		184	29.11%	149	24.67%	199	25.42%	532	26.35%
	No		384	60.76%	394	65.23%	501	63.98%	1,279	63.35%
	Prefer not to say		5	0.79%	2	0.33%	6	0.77%	13	0.64%
General health										
	Excellent	1.97 (0.14)	81	12.82%	83	13.74%	127	16.22%	291	14.41%
	Very good		165	26.11%	174	28.81%	223	28.48%	562	27.84%
	Good		221	34.97%	204	33.77%	254	32.44%	679	33.63%
	Fair		140	22.15%	112	18.54%	140	17.88%	392	19.42%
	Poor		25	3.96%	31	5.13%	39	4.98%	95	4.71%

Health dimensions										
Vision	No difficulty	1.60 (0.20)	339	53.64%	343	56.79%	418	53.38%	1100	54.48%
	Slight difficulty		185	29.27%	176	29.14%	228	29.12%	589	29.17%
	Some difficulty		71	11.23%	59	9.77%	96	12.26%	226	11.19%
	A lot of difficulty		35	5.54%	23	3.81%	37	4.73%	95	4.71%
	Unable		2	0.32%	3	0.50%	4	0.51%	9	0.45%
Hearing	No difficulty	5.39 (0.01)	410	64.87%	444	73.51%	521	66.54%	1375	68.10%
	Slight difficulty		140	22.15%	104	17.22%	166	21.20%	410	20.31%
	Some difficulty		58	9.18%	41	6.79%	77	9.83%	176	8.72%
	A lot of difficulty		19	3.01%	14	2.32%	16	2.04%	49	2.43%
	Unable		5	0.79%	1	0.17%	3	0.38%	9	0.45%
Getting around	No difficulty	1.60 (0.20)	461	72.94%	471	77.98%	572	73.05%	1504	74.49%
	Slight difficulty		102	16.14%	73	12.09%	117	14.94%	292	14.46%
	Some difficulty		44	6.96%	34	5.63%	57	7.28%	135	6.69%
	A lot of difficulty		22	3.48%	25	4.14%	34	4.34%	81	4.01%
	Unable		3	0.47%	1	0.17%	3	0.38%	7	0.35%
Day-to-day activities	No difficulty	2.58 (0.08)	378	59.81%	414	68.54%	492	62.84%	1284	63.60%
	Slight difficulty		151	23.89%	99	16.39%	160	20.43%	410	20.31%
	Some difficulty		63	9.97%	58	9.60%	74	9.45%	195	9.66%
	A lot of difficulty		33	5.22%	29	4.80%	52	6.64%	114	5.65%
	Unable		7	1.11%	4	0.66%	5	0.64%	16	0.79%
Sleep	None of the time	1.36 (0.26)	169	26.74%	166	27.48%	172	21.97%	507	25.11%
	Only occasionally		191	30.22%	186	30.79%	276	35.25%	653	32.34%
	Sometimes		136	21.52%	136	22.52%	157	20.05%	429	21.25%
	Often		86	13.61%	74	12.25%	118	15.07%	278	13.77%

	Most or all of the time		50	7.91%	42	6.95%	60	7.66%	152	7.53%
Exhausted	None of the time	1.83 (0.16)	169	26.74%	160	26.49%	187	23.88%	516	25.56%
	Only occasionally		183	28.96%	185	30.63%	230	29.37%	598	29.62%
	Sometimes		161	25.47%	143	23.68%	194	24.78%	498	24.67%
	Often		83	13.13%	86	14.24%	119	15.20%	288	14.26%
	Most or all of the time		36	5.70%	30	4.97%	53	6.77%	119	5.89%
Lonely	None of the time	4.75 (0.01)	311	49.21%	274	45.36%	305	38.95%	890	44.08%
	Only occasionally		138	21.84%	134	22.19%	193	24.65%	465	23.03%
	Sometimes		102	16.14%	103	17.05%	170	21.71%	375	18.57%
	Often		52	8.23%	54	8.94%	77	9.83%	183	9.06%
	Most or all of the time		29	4.59%	39	6.46%	38	4.85%	106	5.25%
Support	None of the time	3.68 (0.03)	334	52.85%	317	52.48%	365	46.62%	1016	50.32%
	Only occasionally		126	19.94%	124	20.53%	168	21.46%	418	20.70%
	Sometimes		105	16.61%	93	15.40%	142	18.14%	340	16.84%
	Often		50	7.91%	45	7.45%	77	9.83%	172	8.52%
	Most or all of the time		17	2.69%	25	4.14%	31	3.96%	73	3.62%
Concentration/thinking clearly	None of the time	8.73 (<0.01)	293	46.36%	284	47.02%	293	37.42%	870	43.09%
	Only occasionally		153	24.21%	152	25.17%	216	27.59%	521	25.80%
	Sometimes		109	17.25%	107	17.72%	153	19.54%	369	18.28%
	Often		56	8.86%	42	6.95%	83	10.60%	181	8.96%
	Most or all of the time		21	3.32%	19	3.15%	38	4.85%	78	3.86%
Anxious	None of the time	11.75 (<0.01)	255	40.35%	249	41.23%	246	31.42%	750	37.15%
	Only occasionally		165	26.11%	157	25.99%	220	28.10%	542	26.84%
	Sometimes		125	19.78%	110	18.21%	156	19.92%	391	19.37%
	Often		60	9.49%	63	10.43%	100	12.77%	223	11.05%
	Most or all of the time		27	4.27%	25	4.14%	61	7.79%	113	5.60%
Depression	None of the time	6.51	277	43.83%	270	44.70%	276	35.25%	823	40.76%

	Only occasionally	(<0.01)	159	25.16%	158	26.16%	217	27.71%	534	26.45%
	Sometimes		106	16.77%	82	13.58%	152	19.41%	340	16.84%
	Often		61	9.65%	61	10.10%	89	11.37%	211	10.45%
	Most or all of the time		29	4.59%	33	5.46%	49	6.26%	111	5.50%
Control	None of the time	6.85 (<0.01)	323	51.11%	313	51.82%	333	42.53%	969	47.99%
	Only occasionally		127	20.09%	143	23.68%	198	25.29%	468	23.18%
	Sometimes		97	15.35%	90	14.90%	141	18.01%	328	16.25%
	Often		57	9.02%	35	5.79%	69	8.81%	161	7.97%
	Most or all of the time		28	4.43%	23	3.81%	42	5.36%	93	4.61%
Physical Pain	No physical pain	0.20 (0.82)	205	32.44%	204	33.77%	271	34.61%	680	33.68%
	Mild physical pain		272	43.04%	274	45.36%	348	44.44%	894	44.28%
	Moderate physical pain		128	20.25%	87	14.40%	116	14.81%	331	16.39%
	Severe physical pain		23	3.64%	32	5.30%	35	4.47%	90	4.46%
	Very severe physical pain		4	0.63%	7	1.16%	13	1.66%	24	1.19%

[1] Analysis of variance (ANOVA) tested the variance among the three cohorts to test if their baseline health condition was significantly different. The F-statistic, combined with the degree of freedom information, significant with a 5% significance level.

Table E-3: UK sample data quality check, understanding and decision-making

		Cohort 1		Cohort 2		Cohort3		Overall	
		Number	Proportion	Number	Proportion	Number	Proportion	Number	Proportion
		627	100.00%	600	100.00%	776	100.00%	2003	100.00%
Data quality check									
Negative feedback		27	4.31%	24	4.00%	32	4.12%	83	4.14%
Shorter than 10 min		196	31.26%	177	29.50%	160	20.62%	533	26.61%
Question shorter than 10 seconds		100	15.95%	99	16.50%	109	14.05%	308	15.38%
Repeated employment status question inconsistency		17	2.71%	12	2.00%	26	3.35%	55	2.75%
Left-right bias: left		3	0.48%	4	0.67%	11	1.42%	18	0.90%
Left-right bias: right		1	0.16%	1	0.17%	6	0.77%	8	0.40%
Reason for not selecting WTD state (Cohort 3)						79	11.28%		
	In the questions, there were always better options in either Life A or Life B.					60	7.73%		
	Being alive, even with the given health and wellbeing problems, is always better than being dead.					17	2.4%		
	I choose “immediate death” as worst because of my religious beliefs.					6	0.77%		
	I choose “immediate death” as worst because of my outlook on life or family related considerations.					48	6.19%		
	I found it difficult to imagine what “immediate death” would mean so I did not consider it.					9	1.16%		
Reason for not selecting WTD state – because of study design only						18	10.72%		
Reason for not selecting WTD state – because of dead state interpretation difficulty only						4	4.92%		

Understanding									
It was easy to understand the questions I was asked									
	Strongly agree	288	45.93%	289	48.17%	319	41.11%	896	44.73%
	Agree	285	45.45%	274	45.67%	366	47.16%	925	46.18%
	Neither agree nor disagree	41	6.54%	23	3.83%	53	6.83%	117	5.84%
	Disagree	9	1.44%	13	2.17%	35	4.51%	57	2.85%
	Strongly disagree	4	0.64%	1	0.17%	3	0.39%	8	0.40%
I found it easy to tell the difference between the two lives I was asked to think about.									
	Strongly agree	229	36.52%	222	37.00%	254	32.73%	705	35.20%
	Agree	292	46.57%	272	45.33%	355	45.75%	919	45.88%
	Neither agree nor disagree	66	10.53%	64	10.67%	109	14.05%	239	11.93%
	Disagree	35	5.58%	42	7.00%	52	6.70%	129	6.44%
	Strongly disagree	5	0.80%		0.00%	6	0.77%	11	0.55%
I found it difficult to decide on my answers to the questions.									
	Strongly agree	61	9.73%	89	14.83%	103	13.27%	253	12.63%
	Agree	215	34.29%	169	28.17%	267	34.41%	651	32.50%
	Neither agree nor disagree	141	22.49%	131	21.83%	172	22.16%	444	22.17%
	Disagree	163	26.00%	162	27.00%	169	21.78%	494	24.66%
	Strongly disagree	47	7.50%	49	8.17%	65	8.38%	161	8.04%
The number of choice questions I had to answer was appropriate (not too many or too few)									
	Strongly agree	180	28.71%	194	32.33%	226	29.12%	600	29.96%
	Agree	296	47.21%	264	44.00%	349	44.97%	909	45.38%
	Neither agree nor disagree	119	18.98%	93	15.50%	127	16.37%	339	16.92%
	Disagree	24	3.83%	37	6.17%	68	8.76%	129	6.44%
	Strongly disagree	8	1.28%	12	2.00%	6	0.77%	26	1.30%
Amount of information used to describe the health and wellbeing problems was appropriate (not too much or too little)									

	Strongly agree	192	30.62%	217	36.17%	244	31.44%	653	32.60%
	Agree	321	51.20%	285	47.50%	390	50.26%	996	49.73%
	Neither agree nor disagree	94	14.99%	76	12.67%	106	13.66%	276	13.78%
	Disagree	14	2.23%	20	3.33%	30	3.87%	64	3.20%
	Strongly disagree	6	0.96%	2	0.33%	6	0.77%	14	0.70%
Confident with the answers									
	Very confident	138	22.01%	184	30.67%	199	25.64%	521	26.01%
	Confident	378	60.29%	322	53.67%	424	54.64%	1,124	56.12%
	Neither confident nor unconfident (neutral)	94	14.99%	79	13.17%	117	15.08%	290	14.48%
	Unconfident	14	2.23%	14	2.33%	33	4.25%	61	3.05%
	Very unconfident	3	0.48%	1	0.17%	3	0.39%	7	0.35%

Decision making strategies									
	Considered all of the health and wellbeing aspects all the time	173	27.59%	166	27.67%	287	36.98%	626	31.25%
	Only considered health and wellbeing aspects that I believe to be important	229	36.52%	201	33.50%	293	37.76%	723	36.10%
	Considered different health and wellbeing aspects each time	98	15.63%	87	14.50%	139	17.91%	324	16.18%
	Mainly considered the length of time in Life A or B	106	16.91%	120	20.00%	24	3.09%	250	12.48%
	Considered health and wellbeing aspects not presented here	2	0.32%	1	0.17%	5	0.64%	8	0.40%
	Other decision method	11	1.75%	16	2.67%	15	1.93%	42	2.10%
	Select randomly	2	0.32%	3	0.50%	3	0.39%	8	0.40%
	Do not know	6	0.96%	6	1.00%	10	1.29%	22	1.10%
	11	45	7.18%	43	7.17%	56	7.22%	144	7.19%

Table E-4: UK sample data quality check, understanding and decision-making

		Cohort 1		Cohort 2		Cohort3		Overall	
		Number	Proportion	Number	Proportion	Number	Proportion	Number	Proportion
		632	100.00%	604	100.00%	783	100.00%	2019	100.00%
Data quality check									
Negative feedback		15	2.37%	12	1.99%	29	3.70%	56	2.77%
Shorter than 10 min		146	23.10%	121	20.03%	161	20.56%	428	21.20%
Question shorter than 10 seconds		129	20.41%	108	17.88%	145	18.52%	382	18.92%
Repeated employment status question inconsistency		16	2.53%	25	4.14%	30	3.83%	71	3.52%
Left-right bias: left		12	1.90%	9	1.49%	31	3.96%	52	2.58%
Left-right bias: right		1	0.16%	4	0.66%	3	0.38%	8	0.40%
Dominant question failed		104	16.46%	114	18.87%	122	15.58%	340	16.84%
Reason for not selecting WTD state (Cohort 3)						186	23.75%		
	In the questions, there were always better options in either Life A or Life B.					64	8.17%		
	Being alive, even with the given health and wellbeing problems, is always better than being dead.					135	17.24%		
	I choose “immediate death” as worst because of my religious beliefs.					13	1.66%		
	I choose “immediate death” as worst because of my outlook on life or family related considerations.					44	5.62%		
	I found it difficult to imagine what “immediate death” would mean so I did not consider it.					24	3.07%		
Reason for not selecting WTD state – because of study design only						24	3.07%		
Reason for not selecting WTD state – because of dead state interpretation difficulty only						4	0.51%		

Understanding									
It was easy to understand the questions I was asked									
	Strongly agree	269	42.56%	253	41.89%	302	38.57%	824	40.81%
	Agree	299	47.31%	276	45.70%	388	49.55%	963	47.70%
	Neither agree nor disagree	49	7.75%	54	8.94%	71	9.07%	174	8.62%
	Disagree	12	1.90%	18	2.98%	18	2.30%	48	2.38%
	Strongly disagree	3	0.47%	3	0.50%	4	0.51%	10	0.50%
I found it easy to tell the difference between the two lives I was asked to think about.									
	Strongly agree	208	32.91%	193	31.95%	227	28.99%	628	31.10%
	Agree	298	47.15%	291	48.18%	389	49.68%	978	48.44%
	Neither agree nor disagree	85	13.45%	85	14.07%	117	14.94%	287	14.21%
	Disagree	37	5.85%	32	5.30%	47	6.00%	116	5.75%
	Strongly disagree	4	0.63%	3	0.50%	3	0.38%	10	0.50%
I found it difficult to decide on my answers to the questions.									
	Strongly agree	100	15.82%	63	10.43%	127	16.22%	290	14.36%
	Agree	150	23.73%	142	23.51%	225	28.74%	517	25.61%
	Neither agree nor disagree	156	24.68%	177	29.30%	206	26.31%	539	26.70%
	Disagree	173	27.37%	153	25.33%	170	21.71%	496	24.57%
	Strongly disagree	53	8.39%	69	11.42%	55	7.02%	177	8.77%
The number of choice questions I had to answer was appropriate (not too many or too few)									
	Strongly agree	162	25.63%	165	27.32%	198	25.29%	525	26.00%
	Agree	327	51.74%	277	45.86%	352	44.96%	956	47.35%
	Neither agree nor disagree	95	15.03%	110	18.21%	172	21.97%	377	18.67%
	Disagree	38	6.01%	42	6.95%	54	6.90%	134	6.64%
	Strongly disagree	10	1.58%	10	1.66%	7	0.89%	27	1.34%

Amount of information used to describe the health and wellbeing problems was appropriate (not too much or too little)									
	Strongly agree	202	31.96%	180	29.80%	216	27.59%	598	29.62%
	Agree	310	49.05%	303	50.17%	395	50.45%	1,008	49.93%
	Neither agree nor disagree	90	14.24%	94	15.56%	147	18.77%	331	16.39%
	Disagree	22	3.48%	23	3.81%	22	2.81%	67	3.32%
	Strongly disagree	8	1.27%	4	0.66%	3	0.38%	15	0.74%
Confident with the answers									
	Very confident	208	32.91%	220	36.42%	227	28.99%	655	32.44%
	Confident	326	51.58%	289	47.85%	414	52.87%	1,029	50.97%
	Neither confident nor unconfident (neutral)	85	13.45%	78	12.91%	120	15.33%	283	14.02%
	Unconfident	11	1.74%	14	2.32%	19	2.43%	44	2.18%
	Very unconfident	2	0.32%	3	0.50%	3	0.38%	8	0.40%

Decision making strategies									
	Considered all of the health and wellbeing aspects all the time	216	34.18%	193	31.95%	328	41.89%	737	36.50%
	Only considered health and wellbeing aspects that I believe to be important	192	30.38%	201	33.28%	274	34.99%	667	33.04%
	Considered different health and wellbeing aspects each time	85	13.45%	90	14.90%	130	16.60%	305	15.11%
	Mainly considered the length of time in Life A or B	91	14.40%	81	13.41%	15	1.92%	187	9.26%
	Considered health and wellbeing aspects not presented here	9	1.42%	11	1.82%	9	1.15%	29	1.44%
	Other decision method	19	3.01%	15	2.48%	16	2.04%	50	2.48%
	Select randomly	8	1.27%	6	0.99%	4	0.51%	18	0.89%
	Do not know	12	1.90%	7	1.16%	7	0.89%	26	1.29%

Table E-5: UK regression results (multiple models)

Attribute	Model 1	Model 2	Model 3	Model 4	Model 5
Seeing (anchored)					
2	-0.067	-0.059	-0.067	-0.063	-0.067
3	-0.096	-0.082	-0.151	-0.148	-0.096
4	-0.225	-0.175	-0.292	-0.286	-0.225[2]
5	-0.369	-0.340	-0.445	-0.439	-0.369
Hearing (anchored)					
2	-0.046	-0.012**	-0.007**	-0.005**	-0.046
3	-0.061	-0.047	-0.051	-0.048	-0.061
4	-0.097	-0.110	-0.094	-0.086	-0.097
5	-0.120	-0.197	-0.182	-0.179	-0.120[2]
Getting around (anchored)					
2	-0.029*	-0.029**	-0.039	-0.040	-0.029*
3	-0.044	0.003**[1]	-0.043	-0.044	-0.044
4	-0.105	-0.082	-0.093	-0.092	-0.105
5	-0.140	-0.203	-0.210	-0.208	-0.140[2]
Day to day activities (anchored)					
2	-0.018**	-0.016**	-0.043	-0.034	-0.018**
3	-0.062	-0.045	-0.054	-0.051	-0.062

4	-0.112	-0.080	-0.111	-0.108	-0.112
5	-0.168	-0.177	-0.195	-0.186	-0.168
Sleeping (anchored)					
2	-0.020**	0.066[1]	0.012**[1]	0.008**	-0.020**[2]
3	-0.026**	0.031[1]	-0.013**	-0.015**	-0.026**[2]
4	-0.049	-0.010**[1]	-0.024**	-0.023**	-0.049[2]
5	-0.037*[1]	-0.011**[1]	-0.033	-0.032	-0.037*[1]
Exhausted (anchored)					
2	0.008**[1]	-0.013**	0.006**[1]	0.006**	0.008**[1]
3	0.003**[1]	-0.055	-0.027*	-0.027*	0.003**[1]
4	-0.041	-0.084	-0.038	-0.041	-0.041
5	-0.059	-0.030**[1]	-0.076	-0.073	-0.059
Lonely (anchored)					
2	0.001**	-0.009**	-0.003**	-0.004**	0.001**
3	0.003**[1]	-0.005**[1]	-0.021**	-0.020**	0.003**[1]
4	0.001[1]	-0.060	-0.034	-0.035	0.001[12]
5	-0.036*	-0.052[1]	-0.054	-0.051	-0.036*
Unsupported (anchored)					
2	-0.038*	-0.003**	-0.002**	-0.005**	-0.038*
3	-0.050	-0.003**[1]	-0.019**	-0.019**	-0.050
4	-0.087	-0.039	-0.057	-0.055	-0.087
5	-0.029**[1]	-0.032*[1]	-0.058	-0.057	-0.029**[1]

Concentration (anchored)					
2	-0.044	0.012**[1]	-0.024*	-0.029*	-0.044
3	-0.082	-0.040	-0.175	-0.154	-0.082
4	-0.019**[1]	-0.059	-0.046[1]	-0.048	-0.019**[1]
5	-0.055	-0.080	-0.052	-0.057	-0.055
Anxious (anchored)					
2	0.017**[1]	-0.006**	0.001**[1]	-0.003**	0.017**[1]
3	0.018**[1]	-0.020**	-0.009**	-0.017**	0.018**[1]
4	-0.032**	-0.034**	-0.022	0.009	-0.032**
5	-0.038	-0.073	-0.030	-0.030*	-0.038
Depression (anchored)					
2	-0.053	0.017**[1]	-0.029	-0.032	-0.053[2]
3	-0.032**[1]	-0.021**	-0.046	-0.048	-0.032**[1]
4	-0.108	-0.059	-0.043[1]	-0.047	-0.108
5	-0.149	-0.097	-0.091	-0.089	-0.149
Control (anchored)					
2	0.020**[1]	-0.044	-0.018**	-0.017**	0.020**[12]
3	-0.013**	-0.024**[1]	-0.022**	-0.021**	-0.013**
4	-0.011**[1]	-0.032	-0.151	-0.199	-0.011**[1]
5	-0.060	-0.069	-0.065[1]	-0.068	-0.060
Physical pain (anchored)					

2	-0.044	-0.081	-0.021**	-0.025**	-0.044
3	-0.111	-0.144	-0.074	-0.075	-0.111
4	-0.294	-0.283	-0.289	-0.290	-0.294
5	-0.373	-0.441	-0.396	-0.392	-0.373
Duration/Death (unanchored)	0.616	1.099	-2.642	2.667	0.616
Hearing level 5 (order term)					-0.102
Getting around level 5 (order term)					-0.063
Sleep level 2 (order term)					0.054
Sleep level 4 (order term)					0.038
Depression level 2 (order term)					0.072
Control level 2 (order term)					-0.061
Order main effect					-0.054**
Observations	16,218	15,548	60,528	30,264	31,688
Log-likelihood	-4544.945	-4393.045	-18443.209	-15970.090	-8937.991
Pseudo R-squared	0.1914	0.181	0.121	0.071	0.1861
Logical inconsistent	10	14	4	4	10
Insignificant 10%	17	19	17	17	17
Utility of health state with all items at level 2	0.686	0.822	0.764	0.758	0.686
Utility of health state with all items at level 3	0.448	0.556	0.453	0.452	0.448
Utility of health state with all items at level 4	-0.179	-0.083	-0.133	-0.130	-0.179

Utility of worst state	-0.634	-0.791	-0.887	-0.862	-0.634
Proportion of WTD[3] states	17.16%	16.74%	27.36%		

* The level is not significant at 5%; ** the level is not significant at 10%

[1] the attribute level is logically inconsistent with the previous one.

[2] the attribute level that the interaction term is significant.

[3] WTD: worse than dead state

Model explanation: Model 1: conditional logit regression with standard-order DCE_{TTO}; Model 2 conditional logit regression with revised-order DCE_{TTO}; Model 3: conditional logit regression with DCE-Death; Model 4: rank ordered logit model with DCE-Death; Model 5: conditional logit regression with pooled DCE_{TTO} data, including both standard-order DCE_{TTO} and revised-order DCE_{TTO}

Table E-6: Australian data regression results (multiple models)

Attribute	Model 1	Model 2	Model 3	Model 4	Model 5
Seeing (anchored)					
2	-0.037*	-0.081	-0.071	-0.073	-0.037*
3	-0.039**	-0.087	-0.107	-0.109	-0.039**[1]
4	-0.161	-0.148	-0.236	-0.230	-0.161
5	-0.308	-0.258	-0.366	-0.366	-0.308
Hearing (anchored)					
2	-0.066	-0.064	-0.009**	-0.011**[1]	-0.066
3	-0.083*	-0.082	-0.047	-0.039	-0.083*
4	-0.101	-0.103	-0.064	-0.065	-0.101
5	-0.139	-0.179	-0.138	-0.135	-0.139
Getting around (anchored)					
2	-0.003**	-0.064	-0.048	-0.039	-0.003**
3	-0.015**	-0.050[1]	-0.067	-0.062	-0.015**
4	-0.133	-0.085	-0.095	-0.091	-0.133
5	-0.134	-0.203	-0.167	-0.164	-0.134
Day to day activities (anchored)					
2	-0.064	-0.046	-0.023*	-0.030*	-0.064

3	-0.033**[1]	-0.059	-0.050	-0.053	-0.033**[1]
4	-0.101	-0.103	-0.095	-0.097	-0.101
5	-0.164	-0.143	-0.174	-0.181	-0.164
Sleeping (anchored)					
2	0.002**[1]	0.051**[1]	-0.010**	-0.006**	0.002**[1]
3	0.010**[1]	0.043*[1]	-0.026*	-0.023**	0.010**[1]
4	-0.021**	0.023**[1]	-0.048	-0.045	-0.021**
5	-0.021**	-0.027**	-0.057	-0.049	-0.021**
Exhausted (anchored)					
2	-0.007**	-0.020**	-0.015**	-0.010**	-0.007**
3	-0.007**[1]	-0.060	-0.028*	-0.026**	-0.007**[1]
4	-0.024**	-0.048[1]	-0.076	-0.076	-0.024**
5	-0.028**	-0.055	-0.090	-0.088	-0.028**[1]
Lonely (anchored)					
2	-0.010**	0.021**[1]	0.011**	0.010**	-0.010**
3	-0.010**[1]	-0.005**	-0.021**	-0.020**	-0.010**[1]
4	-0.051*	-0.043*	-0.042	-0.039	-0.051*
5	-0.051*	-0.009**[1]	-0.049	-0.053	-0.051*
Unsupported (anchored)					
2	-0.032**	-0.055	0.007**[1]	0.011**[1]	-0.032**
3	-0.053**	-0.028**[1]	-0.009**	-0.005**	-0.053**[1]
4	-0.088**	-0.075	-0.038	-0.026**	-0.088**

5	-0.043**[1]	-0.057[1]	-0.044	-0.036	-0.043**[1]
Concentration (anchored)					
2	0.006**[1]	-0.017	-0.024**	-0.021**	0.006**[1]
3	-0.031**	-0.022	-0.046	-0.037	-0.031**
4	0.007**[1]	-0.061	-0.049	-0.047	0.007**[1]
5	-0.025**	-0.030**[1]	-0.066	-0.064	-0.025**
Anxious (anchored)					
2	-0.017**	0.009**[1]	-0.030	-0.009	-0.017**
3	-0.015**[1]	-0.012**	-0.016**[1]	-0.016**	-0.015**[1]
4	-0.031**	-0.052	-0.045	-0.042	-0.031**
5	-0.060	-0.042[1]	-0.046	-0.046	-0.060
Depression (anchored)					
2	-0.014**	-0.006**	-0.020**	-0.021**	-0.014**
3	0.006**[1]	-0.008**	-0.031	-0.032*	0.006**[1]
4	-0.054*	-0.041*	-0.048	-0.043	-0.054*
5	-0.116	-0.052	-0.078	-0.070	-0.116
Control (anchored)					
2	0.008**[1]	-0.048	-0.036	-0.033	0.008**[1]
3	-0.018**	-0.006**[1]	-0.003**[1]	-0.012**[1]	-0.018**
4	-0.028**	-0.039	-0.035	-0.039	-0.028**
5	-0.065	-0.079	-0.064	-0.069	-0.065

Physical pain (anchored)					
2	-0.068	-0.081	-0.059	-0.058	-0.068
3	-0.112	-0.136	-0.076	-0.082	-0.112
4	-0.304	-0.250	-0.275	-0.285	-0.304
5	-0.433	-0.405	-0.374	-0.378	-0.433
Duration/Death (unanchored)	0.451	0.548	-2.462	-2.462	0.451
Hearing level 5 (order term)					-0.026
Getting around level 1 (order term)					-0.034
Getting around level 5 (order term)					-0.051
Concentration level 4 (order term)					-0.037
Order main effect					0.099
Observations	16458	15600	61074	30498	32110
Log-likelihood	-4919.716	-4640.015	-18772.880	-16420.360	-9559.731
Pseudo R-squared	0.139	0.143	0.097	0.062	0.141
Logical inconsistent (summarize it)	8	11	4	2	8
Insignificant 10%	26	15	11	14	26
Utility of health state with all items at level 2	0.699	0.598	0.673	0.711	0.699
Utility of health state with all items at level 3	0.599	0.488	0.472	0.483	0.599
Utility of health state with all items at level 4	-0.089	-0.025	-0.146	-0.126	-0.089

Utility of worst state	-0.588	-0.539	-0.713	-0.699	-0.588
Proportion of WTD[3] states	13.16%	12.75%	17.05%		

* The level is not significant at 5%; ** the level is not significant at 10%

[1] the attribute level is logically inconsistent with the previous one.

[2] the attribute level that the interaction term is significant.

[3] WTD: worse than dead state

Model explanation: Model 1: conditional logit regression with standard-order DCE_{TTO}; Model 2 conditional logit regression with revised-order DCE_{TTO}; Model 3: conditional logit regression with DCE-Death; Model 4: rank ordered logit model with DCE-Death; Model 5: conditional logit regression with pooled DCE_{TTO} data, including both standard-order DCE_{TTO} and revised-order DCE_{TTO}

Table E-7: Non-significant levels, positive disutility and disordered levels

a. By attributes

			In total	Vision	Hearing	Mobility	Daily activity	Slee p	Fatigu e	Lonelines s	Suppor t	Concentratin g	Contr ol	Anxious	depress	Pain severity
UK	Model 1	non-significant	17	0	0	0	1	2	2	3	1	1	3	3	1	0
		disordered	7	0	0	0	0	1	0	1	1	1	1	1	1	0
		incorrect sign	8	0	0	0	0	0	2	3	0	0	1	2	0	0
		combined	1	0	0	0	0	1	0	0	0	0	0	0	0	0
	Model 2	non-significant	20	0	1	2	1	3	2	2	2	1	1	3	2	0
		disordered	6	0	0	1	0	0	1	2	1	0	1	0	0	0
		incorrect sign	5	0	0	0	0	3	0	0	0	1	0	0	1	0
		combined	3	0	0	0	0	1	0	1	1	0	0	0	0	0
	Model 3	non-significant	18	0	1	0	0	3	1	2	2	1	3	3	1	1
		disordered	4	0	0	0	0	0	0	0	0	1	1	1	1	0
		incorrect sign	4	0	0	0	0	1	1	0	0	0	0	2	0	0
		combined	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUS	Model 1	non-significant	26	1	0	2	1	4	4	2	0	4	3	3	2	0
		disordered	6	0	0	0	1	1	0	0	1	1	0	1	1	0
		incorrect sign	6	0	0	0	0	2	0	0	0	2	1	0	1	0
		combined	1	0	0	0	0	0	0	0	1	0	0	0	0	0
	Model 2	non-significant	16	0	0	0	0	3	1	3	1	3	1	2	2	0
		disordered	8	0	0	1	0	0	1	1	2	1	1	1	0	0
		incorrect sign	5	0	0	0	0	3	0	1	0	0	0	1	0	0
		combined	5	0	0	1	0	1	1	0	1	0	0	1	0	0
	Model 3	non-significant	11	0	1	0	0	1	1	2	2	1	1	1	1	0
		disordered	2	0	0	0	0	0	0	0	0	0	1	1	0	0

		incorrect sign	2	0	0	0	0	0	0	1	1	0	0	0	0	0
		combined	0	0	0	0	0	0	0	0	0	0	0	0	0	0

b. By levels

			Level 2			Level 3			Level 4			Level 5		
			Total	Severity levels	Frequency levels	Total	Severity levels	Frequency levels	Total	Severity levels	Frequency levels	Total	Severity levels	Frequency levels
UK	Model 1	non-significant	6	1	5	6	0	6	4	0	4	1	0	1
		disordered	0	0	0	3	0	3	2	0	2	2	0	2
		incorrect sign	4	0	4	3	0	3	1	0	1	0	0	0
		combined	0	0	0	0	0	0	0	0	0	1	0	1
	Model 2	non-significant	9	3	6	7	1	6	2	0	2	2	0	2
		disordered	0	0	0	3	1	2	0	0	0	3	0	3
		incorrect sign	3	0	3	1	0	1	1	0	1	0	0	0
		combined	1	0	1	0	0	0	0	0	0	2	0	2
	Model 3	non-significant	8	2	6	6	0	6	4	0	4	0	0	0
		disordered	0	0	0	1	0	1	3	0	3	0	0	0
		incorrect sign	3	0	3	0	0	0	1	0	1	0	0	0
		combined	0	0	0	0	0	0	0	0	0	0	0	0
AUS	Model 1	non-significant	8	1	7	10	3	7	5	0	5	3	0	3
		disordered	0	0	0	4	1	3	1	0	1	1	0	1
		incorrect sign	3	0	3	2	0	2	1	0	1	0	0	0
		combined	0	0	0	0	0	0	0	0	0	1	0	1
	Model 2	non-significant	6	0	6	6	0	6	1	0	1	3	0	3
		disordered	0	0	0	3	1	2	1	0	1	4	0	4
		incorrect sign	3	0	3	1	0	1	1	0	1	0	0	0
		combined	0	0	0	2	1	1	1	0	1	2	0	2
	Model 3	non-significant	6	1	5	5	0	5	0	0	0	0	0	0

		disordered	0	0	0	2	0	2	0	0	0	0	0	0
		incorrect sign	2	0	2	0	0	0	0	0	0	0	0	0
		combined	0	0	0	0	0	0	0	0	0	0	0	0

Highlighted in bold: The Disordered and incorrect sign means that the regression coefficient disutility magnitude of a lower level is smaller than that of a higher level, or has a positive sign for the disutility. The combined are the significant regression coefficients that has a positive sign for the disutility, or the disutility magnitude of a lower level is smaller than that of a higher level.

Non-significance level: 10%.

[1]numbers in each table: sum of the number of non-significant levels, positive disutility and disordered coefficients in three designs. The total number of coefficients for each attribute is 12. The total number of coefficients for each level is 39.

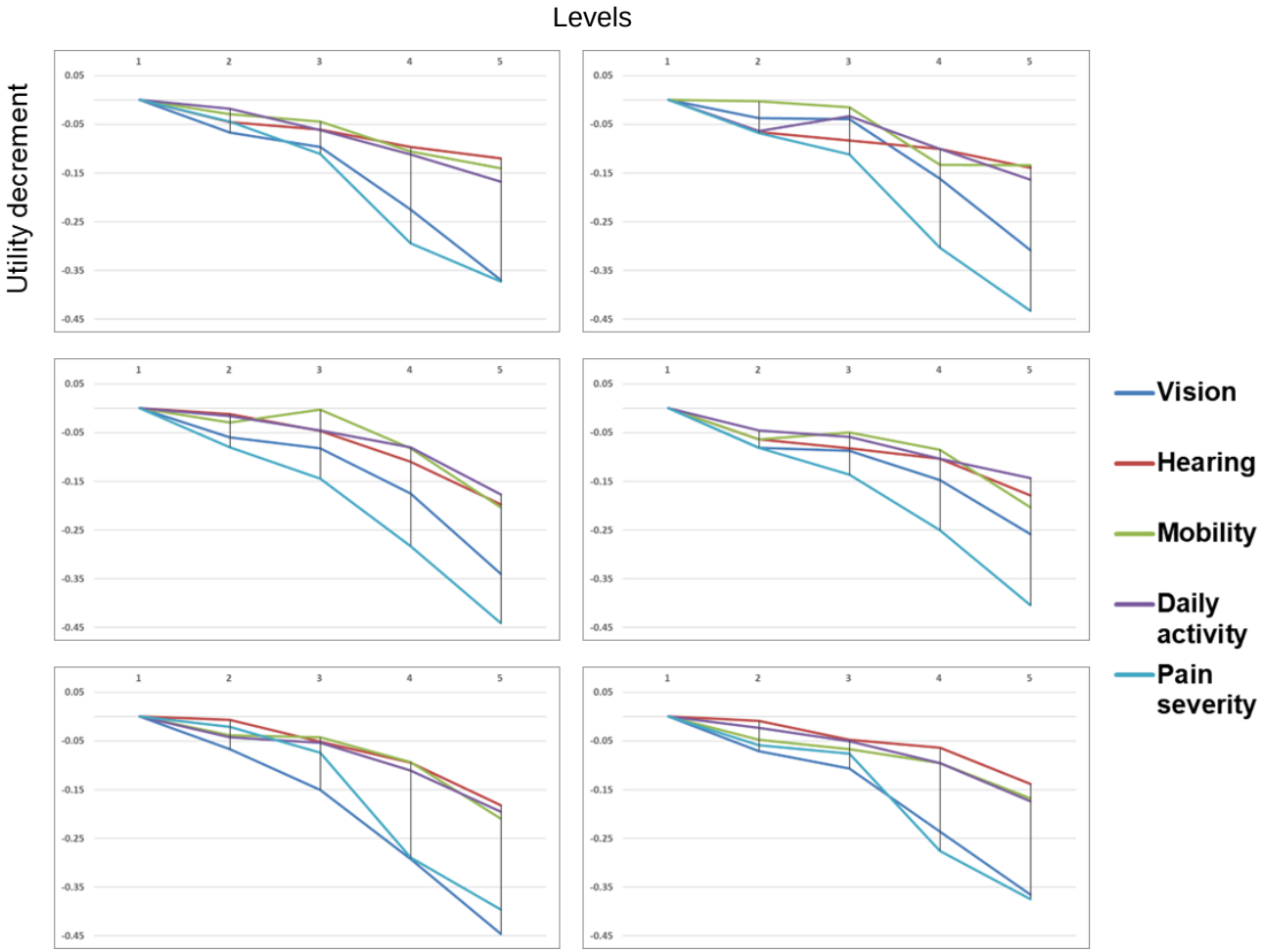
[2]Severity levels: HWB dimension attributes with health problems described by its severity (No difficulty, Slight difficulty, Some difficulty, A lot of difficulty, Unable). Attributes are Vision, Hearing, Mobility, Daily activity and Pain severity.

[3]Frequency levels: HWB dimension attributes with health problems described by its frequency (None of the time, Only occasionally, Sometimes, Often, Most or all of the time).

Attributes are Sleep, Fatigue, Loneliness, Support, Concentrating, Anxious, Sad/depress and Control.

Figure E-1: UK (left) and Australian (right) disutility value by level and dimensions, by models (from top to bottom: Model 1, Model 2, Model 3)

a. Physical health attributes: Vision, Hearing, Mobility, Daily activity and Pain severity



b. Mental health and wellbeing attributes: Sleep, Fatigue, Loneliness, Support, Concentrating, Anxious, Sad/depress and Control

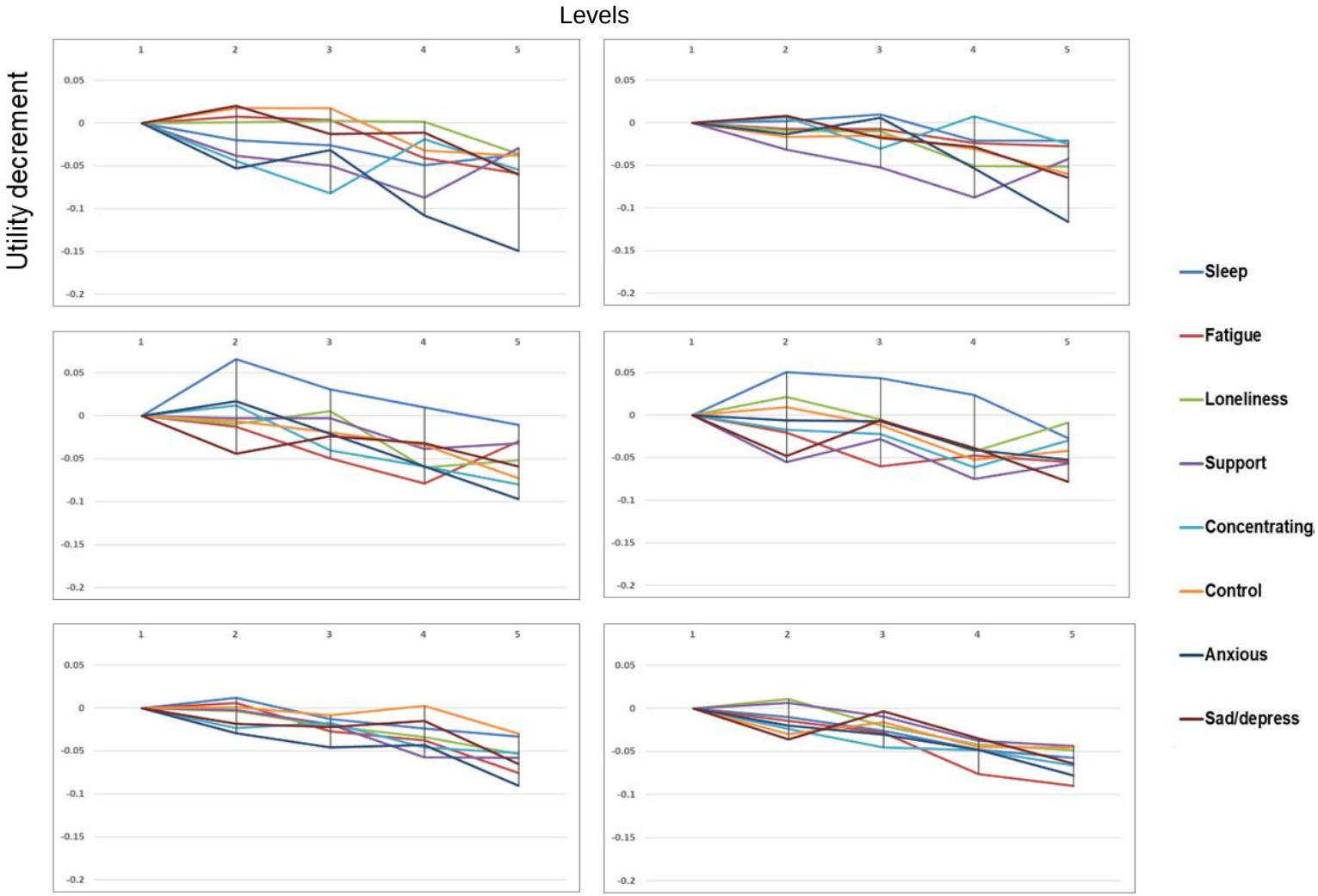
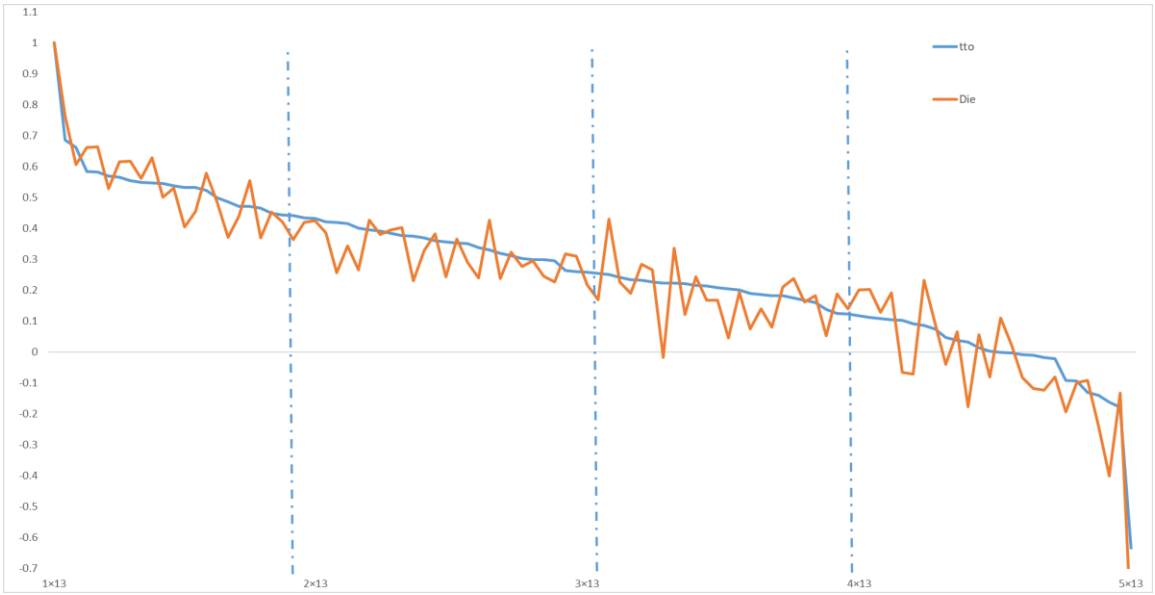
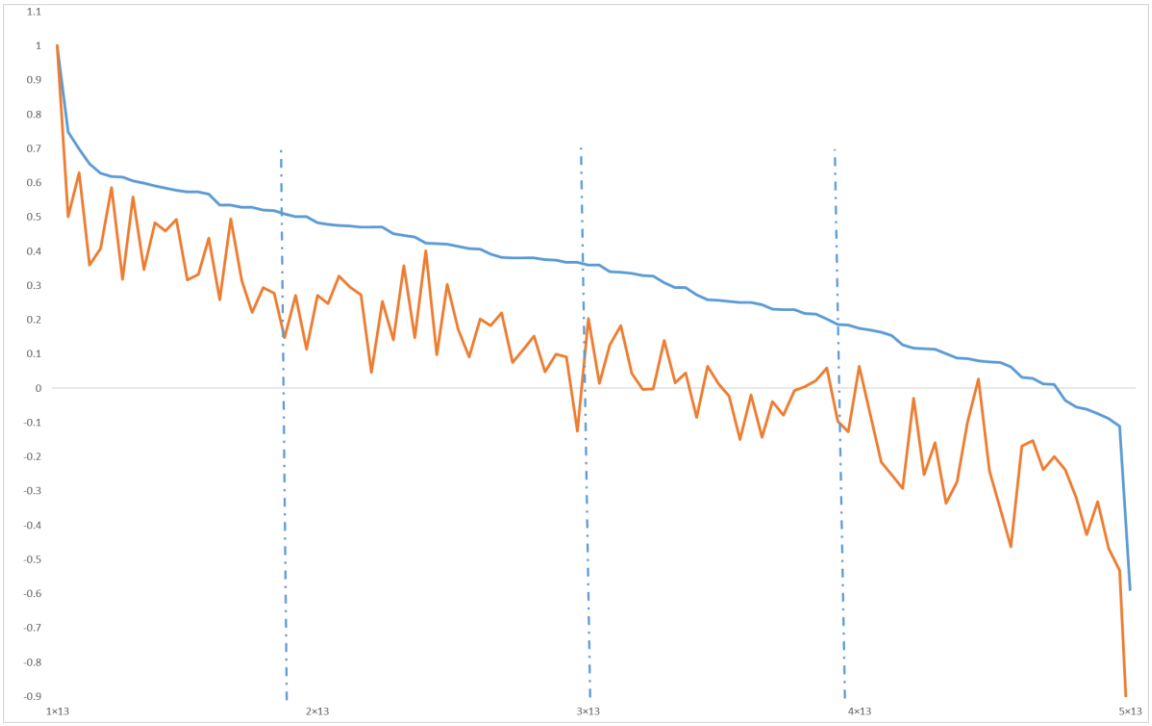


Figure E-2. UK and Australia value sets per selected health state

a: with UK sample



b: with Australian sample



Appendix F: Preference Heterogeneity analysis result, by model

Table F-1: UK MNL regression with correlation terms

Attribute	Gender	Age	Carer or not	Cared or not	General health
Seeing					
2	0.01**	0.02**	-0.01**	0.03**	0.03**
3	0.03**	0.01**	-0.05**	0.05**	0.05**
4	-0.03**	-0.03**	-0.01**	-0.04**	-0.04**
5	-0.03**	-0.12	0.00**	-0.08	-0.08
Hearing					
2	0.06	-0.05**	0.08	0.01**	0.01**
3	0.05**	-0.06**	0.05**	-0.01**	-0.01**
4	-0.01**	-0.08	0.10	0.00**	0.00**
5	0.00**	-0.07*	0.03**	-0.03**	-0.03**
Getting around					
2	0.00**	-0.05**	0.03**	-0.02**	-0.02**
3	-0.02**	-0.05**	-0.01**	0.04**	0.04**
4	-0.03**	-0.06**	0.04**	0.04**	0.04**
5	-0.05	-0.09	0.06*	0.04**	0.04**
Day to day activities					
2	0.04**	-0.05**	0.02**	0.02**	0.02**
3	0.01**	-0.01**	-0.01**	0.03**	0.03**

4	0.01**	-0.03**	0.04**	0.05*	0.05*
5	-0.01**	-0.06**	0.00**	0.04**	0.04**
Sleeping					
2	0.01**	0.02**	-0.01**	0.06**	0.07
3	0.00**	0.07**	-0.01**	0.09*	0.03**
4	-0.02**	0.04**	-0.08	-0.02**	0.04**
5	-0.04**	0.07*	-0.01**	0.05**	0.05**
Exhausted					
2	0.01**	-0.02**	-0.04**	-0.04**	0.00**
3	-0.01**	0.03**	0.01**	0.00**	0.00**
4	0.00**	-0.05**	0.03**	0.06**	0.00**
5	-0.03**	0.00**	0.01**	0.05**	0.01**
Lonely					
2	-0.06**	-0.03**	0.01**	0.00**	0.00**
3	-0.05	0.03**	0.03**	-0.05**	0.01**
4	-0.07**	-0.01**	0.02**	0.00**	-0.03**
5	-0.09**	0.00**	0.04**	0.05**	-0.01**
Unsupported					
2	-0.02**	-0.04**	0.03**	0.07**	-0.02**
3	-0.02**	0.00**	0.02**	0.04**	-0.02**
4	-0.02**	-0.05**	0.04**	0.04**	-0.05**
5	-0.01**	0.00**	0.05**	-0.01**	0.05*

Concentration					
2	0.01**	-0.03**	-0.02**	0.00**	0.03**
3	0.05**	-0.04**	-0.04**	-0.10	0.00**
4	0.02*	-0.03**	-0.01**	-0.02**	-0.01**
5	0.00**	-0.09	-0.04**	-0.07**	-0.02**
Anxious					
2	0.03**	0.04**	0.04**	0.04**	0.01**
3	0.03**	0.03**	0.03**	0.07**	-0.01**
4	-0.05**	-0.03**	0.08*	0.10*	-0.01**
5	-0.03**	0.03**	0.07*	0.08**	0.00**
Depression					
2	0.03**	0.00**	-0.06*	-0.03**	-0.02**
3	0.05**	-0.05**	0.04**	-0.05**	0.00**
4	0.03*	-0.01**	-0.03**	-0.05**	-0.02**
5	-0.01**	-0.01**	-0.01**	0.03**	0.00**
Control					
2	-0.08	-0.03**	0.01**	0.05**	-0.04**
3	-0.05	-0.04**	0.04**	0.02**	0.00**
4	-0.01**	-0.05**	0.03**	0.00**	-0.01**
5	-0.04*	-0.10	0.08	0.04**	-0.02**
Physical pain					

2	-0.01**	-0.01**	0.01**	0.00**	-0.04**
3	-0.06	-0.01**	0.05**	0.03**	-0.04**
4	-0.04**	-0.08	0.04**	0.02**	-0.07
5	-0.10	-0.14	0.09	0.13	-0.07
Time effect (unanchored)	-0.04**	0.44	-0.17**	-0.36**	0.01**
Observation	31850	31850	31850	31850	31850
Log-likelihood	-9033.46	-9036.46	-9072.63	-9065.85	-9071.06
Pseudo R-squared	0.18	0.18	0.18	0.18	0.18

Table F-2: Australian MNL regression with correlation terms

Attribute	Gender	Age	Carer or not	Cared or not	General health
Seeing					
2	-0.04**	0.05**	0.05**	0.00**	0.05**
3	-0.05**	-0.08**	0.11	0.13	0.01**
4	-0.15**	-0.08**	0.22	0.17	-0.04**
5	0.00**	0.00	0.00	0.00	0.00
Hearing					
2	-0.10	0.02**	0.01**	0.10	-0.01**
3	-0.09**	-0.02**	0.02**	0.03**	-0.02**
4	-0.07**	-0.02**	0.07	0.09	0.03**
5	0.00**	0.00	0.00	0.00	0.00
Getting around					
2	-0.06**	0.03**	-0.01**	0.04**	0.01
3	0.00**	-0.12*	-0.01**	0.05**	0.02
4	-0.02**	-0.10**	0.09	0.13	0.03
5	0.00	0.00	0.00	0.00	0.00
Day to day activities					
2	-0.02**	-0.04**	-0.04**	-0.01**	0.04**
3	-0.09**	-0.06**	0.01**	0.05**	-0.03**

4	-0.07**	0.03**	0.08	0.08	0.03**
5	0.00**	0.00	0.00	0.00	0.00
Sleeping					
2	-0.02**	-0.06**	0.03**	0.00**	0.01**
3	0.05**	-0.12**	0.04**	0.03**	0.03**
4	-0.01**	-0.13*	0.02**	-0.03**	0.00**
5	0.00**	0.00	0.00	0.00	0.00
Exhausted					
2	0.09**	0.05**	0.00**	-0.04**	-0.03**
3	0.01**	0.03**	0.03**	0.05**	-0.03**
4	0.08**	-0.03**	0.01**	0.03**	0.01**
5	0.00**	0.00	0.00	0.00	0.00
Lonely					
2	0.06	0.05**	0.03**	0.09	0.07*
3	0.00**	-0.02**	0.11	0.12	0.04**
4	-0.04	0.02**	0.05**	0.15	-0.02**
5	0.00	0.00	0.00	0.00	0.00
Unsupported					
2	0.07**	-0.11**	0.02**	-0.03**	0.04**
3	-0.09**	-0.28	0.02**	0.01**	0.03**
4	-0.07**	-0.25	0.02**	0.05**	0.03**
5	0.00**	0.00	0.00	0.00	0.00

Concentration					
2	0.00**	-0.11**	0.04**	-0.05**	-0.03**
3	0.02*	-0.13*	0.07	-0.03**	-0.01**
4	0.04**	-0.14	0.04**	0.02**	0.01**
5	0.00**	0.00	0.00	0.00	0.00
Anxious					
2	-0.06**	-0.14*	0.04**	-0.02**	0.03**
3	-0.01**	-0.13**	-0.01**	-0.03**	0.00**
4	-0.06**	-0.03**	0.03**	0.02**	-0.04**
5	0.00**	0.00	0.00	0.00	0.00
Depression					
2	-0.03**	-0.03**	0.02**	0.08**	-0.03**
3	-0.04*	0.01**	0.03**	0.04**	-0.01**
4	-0.14**	0.01**	0.05**	0.08*	0.00**
5	0.00**	0.00	0.00	0.00	0.00
Control					
2	0.04	0.06**	0.01**	0.05*	-0.06**
3	-0.02	-0.06**	-0.02**	0.05**	-0.01**
4	0.01**	-0.01**	0.03**	0.06	0.03**
5	0.00	0.00	0.00	0.00	0.00
Physical pain					

2	-0.04**	-0.06**	0.07	0.05**	-0.01**
3	-0.13	-0.17	0.11	0.16**	-0.08
4	-0.10**	-0.26	0.17	0.16**	-0.09
5	0.00	0.00	0.00	0.00	0.00
Time effect (unanchored)	0.05**	0.29	-0.21	-0.25	-0.02**
Observation	31902	31902	31902	31902	31902
Log-likelihood	-9043.49	-9249.43	-9279.65	-9065.85	-9073.06
Pseudo R-squared	0.18	0.22	0.16	0.18	0.18

Table F-3: Latent classes log-likelihood, CAIC and BIC value for various classes for UK and Australian sample

a. UK sample latent classes

Classes	log-likelihood ratio	CAIC	BIC
2	-9280.0	18559.9	18559.9
3	-9162.9	18325.9	18325.9
4	-8698.4	17396.7	17396.7
5	-8998.6	17997.2	17997.2
6	-8942.8	17885.6	17885.6
7	-8854.0	17708.1	17708.1
8	-8761.8	17523.7	17523.7
9	-9086.7	18173.4	18173.4
10	-8936.4	17872.8	17872.8

b. Australian sample latent classes

Classes	log-likelihood ratio	CAIC	BIC
2	-9670.558	19341.12	19341.12
3	-9546.082	19092.16	19092.16
4	-9411.207	18822.41	18822.41
5	convergence not achieved		
6	-8956.883	17913.77	17913.77
7	-9276.494	18552.99	18552.99
8	-9336.979	18673.96	18673.96
9	-9114.141	18228.28	18228.28
10	-9105.633	18211.27	18211.27

Table F-4: Latent class analysis UK and Australian sample

	UK sample				**Australian sample**					
Observation	31902				32032					
Loglikelihood statistic	-9072.621				-9437.916					
Classes	Class I	Class II	Class III	Class IV	Class I	Class II	Class III	Class IV	Class V	Class VI
Seeing										
2	-0.182	-0.056	-1.474	-0.168	-0.745	-0.048	-0.446	0.326	0.058	-0.142
3	-0.328	-0.301	-3.098	-0.176	-0.597	-0.525	-0.393	-0.625	0.203	-0.239
4	-0.453	-0.481	-6.184	-0.495	-0.773	-0.856	-1.251	0.529	-0.031	-0.425
5	-0.417	-1.7	-9.23	-1.277	-0.046	-0.935	-2.566	-0.318	-0.051	-1.041
Hearing										
2	-0.439	0.292	0.34	-0.118	-1.195	-0.106	-0.355	0.629	-0.335	0.152
3	-0.248	-0.436	0.297	-0.075	-0.48	-0.68	-0.219	-0.457	-0.158	0.168
4	-0.132	-0.392	-0.841	-0.377	-0.91	0.332	-0.474	-0.825	-0.231	-0.007
5	-0.278	-0.477	-3.053	-0.63	-0.733	-0.711	-1.159	-1.411	-0.203	-0.182
Getting around										
2	-0.083	-0.288	-0.528	-0.069	0.173	-0.794	-0.012	-0.282	0.081	-0.262
3	-0.016	-0.163	-0.504	-0.258	0.331	0.194	-0.376	-0.767	-0.237	0.076
4	-0.117	-0.932	-2.124	-0.364	0.002	-0.119	-0.683	-0.882	-0.145	-0.434
5	-0.109	-1.116	-1.703	-0.922	0.656	-0.947	-1.208	-2.65	-0.13	-0.509
Day to day activities										
2	-0.069	-0.889	-0.438	0.115	-0.027	0.106	-0.001	-0.73	-0.265	-0.253
3	-0.114	-0.972	-1.06	-0.096	-0.859	0.861	-0.179	-0.398	-0.332	-0.145
4	-0.103	-2.198	-1.351	-0.099	-0.634	-0.286	-0.544	-0.523	-0.21	-0.343
5	-0.186	-2.463	-2.384	-0.541	-0.66	-0.133	-0.979	-1.108	-0.173	-0.574
Sleeping										

2	0.066	-0.343	-0.471	0.268	0.527	0.446	0.107	-0.248	-0.099	0.078
3	-0.002	0.454	0.312	-0.17	-0.214	-0.067	-0.09	0.244	-0.074	0.2
4	-0.104	-0.121	-0.14	-0.169	-0.092	-0.258	-0.269	-0.263	-0.173	0.09
5	-0.029	-0.602	0.015	-0.204	0.065	-0.15	-0.295	-0.33	-0.191	-0.027
Exhausted										
2	-0.118	-0.143	0.18	0.143	-0.84	-0.396	0.227	0.496	-0.151	0.026
3	0.003	-0.574	0.2	0.03	-0.59	0.134	0.137	0.018	-0.073	-0.242
4	-0.139	-0.432	-0.409	-0.147	-1.489	-0.752	-0.013	0.678	-0.021	0.044
5	-0.087	-0.778	-0.619	-0.146	-1.068	-0.07	-0.413	-0.182	0.01	-0.085
Lonely										
2	-0.143	0.297	0.734	0.016	0.259	0.505	-0.332	0.167	-0.029	0.193
3	-0.134	-0.241	0.84	-0.103	0.349	-0.37	-0.237	0.006	0.013	-0.082
4	-0.134	0.029	0.578	-0.382	-0.595	-0.198	-0.665	0.387	0.169	-0.096
5	-0.073	-0.649	0.621	-0.344	-0.776	-1.23	-0.481	1.101	0.374	-0.164
Unsupported (anchored)										
2	0.033	0.032	-0.084	-0.027	0.02	-0.249	-0.027	0.484	0.007	-0.029
3	-0.059	0.574	0.49	-0.11	-0.305	-0.08	-0.079	-0.13	0.11	-0.091
4	0.055	-0.081	0.181	-0.281	0.062	0.4	-0.384	-0.872	-0.102	-0.127
5	-0.029	-0.093	-0.169	-0.276	0.176	-0.021	-0.453	-0.892	-0.082	-0.031
Concentration										
2	0.094	-0.351	-0.531	-0.149	0.117	-0.063	-0.034	-1.093	-0.065	-0.113
3	-0.108	-0.302	-0.633	-0.182	-0.483	0.41	-0.278	0.191	0.069	-0.193
4	-0.208	0.303	-0.884	-0.37	-0.433	0.055	-0.345	-1.087	0.09	-0.119
5	-0.307	-0.77	-1.142	-0.186	-0.453	-0.678	-0.531	-0.86	0.05	-0.029
Anxious										
2	0.021	0.062	0.289	-0.094	-1.216	-0.209	0.199	-1.453	0.331	0.001

3	0.06	0.312	-0.221	-0.242	-0.787	-0.163	0.07	-1.356	0.128	-0.051
4	0.18	-0.086	-0.11	-0.344	-1.043	-0.656	-0.121	-1.115	0.08	-0.004
5	-0.048	-0.742	0.176	-0.35	-0.558	-0.537	-0.415	-1.826	0.006	0.029
Depression										
2	-0.031	0.48	-0.573	-0.355	-0.241	-0.649	-0.334	1.134	0	0.004
3	-0.156	0.487	-0.338	-0.356	-0.04	-0.182	-0.594	-0.163	0.176	-0.047
4	-0.14	-1.065	-0.79	-0.393	-0.149	-0.765	-0.53	-1.096	0.288	-0.221
5	0	-0.394	-1.146	-0.849	-0.196	-0.412	-0.971	-1.566	0.191	-0.064
Control										
2	0.013	-0.101	0.202	0.036	0.07	-0.403	0.152	-0.813	0.079	-0.061
3	0.084	-0.384	0.145	-0.158	-0.436	-0.668	0.043	-0.498	0.131	-0.152
4	0.045	-0.298	0.412	-0.076	-0.17	-1.397	-0.065	-1.624	-0.018	0.003
5	0.021	-1.189	-0.179	-0.08	-0.085	-1.192	-0.355	-1.318	-0.09	-0.125
Physical pain										
2	-0.075	0.63	-0.962	-0.573	-1.629	1.055	-0.438	-1.429	-0.158	-0.107
3	-0.064	-0.035	-1.764	-0.987	-2.387	0.674	-0.686	-3.318	-0.01	-0.566
4	-0.318	-1.412	-2.602	-2.147	-3.75	0.61	-1.747	-5.9	-0.067	-1.234
5	-0.256	-2.008	-3.33	-3.239	-6.064	0.257	-2.276	-9.144	-0.279	-1.701
Class Share	0.27	0.154	0.121	0.455	0.074	0.057	0.257	0.064	0.205	0.343
age	-0.674	0.225	-0.04	0	0.087	0.059	0.137	-0.19	-0.631	0
gender	-0.959	0.104	-0.546	0	-0.149	-0.337	0.386	1.028	-0.175	0
carer	0.553	-0.036	-0.076	0	0.434	0.621	-0.736	-0.47	0.597	0
Cared for	1.247	-0.006	-0.112	0	-0.781	-0.957	-0.53	-0.822	0.345	0
Poor health	0.117	-0.763	0.365	0	-0.063	-0.388	-0.133	0.254	-0.369	0
_cons	1.786	-1.502	-0.519	0	-1.478	-1.382	-0.922	-2.89	0.574	0

Reference

1. Luo, N., et al., *The effects of lead time and visual aids in TTO valuation: a study of the EQ-VT framework.* The European Journal of Health Economics, 2013. **14**: p. 15-24.
2. Jung, J. and C. Tran, *Market inefficiency, insurance mandate and welfare: US health care reform 2010.* Review of Economic Dynamics, 2016. **20**: p. 132-159.
3. Hsiao, W.C., *Abnormal economics in the health sector.* Health policy, 1995. **32**(1-3): p. 125-139.
4. Saloman, J.A., et al., *Measuring and valuing health benefits for economic evaluation*. First edition. ed, ed. J.A. Saloman, et al. 2017, Oxford, United Kingdom: Oxford University Press.
5. Carson, R.T., et al., *Experimental analysis of choice.* Marketing letters, 1994. **5**(4): p. 351-367.
6. Cowles, E., et al., *A review of NICE methods and processes across health technology assessment programmes: why the differences and what is the impact?* Applied health economics and health policy, 2017. **15**(4): p. 469-477.
7. National Institute for, H. and E. Care, *NICE Process and Methods Guides*, in *Guide to the Methods of Technology Appraisal 2013*. 2013, National Institute for Health and Care Excellence (NICE) Copyright © 2013 National Institute for Health and Clinical Excellence, unless otherwise stated. All rights reserved.: London.
8. Brazier, J., et al., *A Review of Generic Preference-Based Measures for Use in Cost-Effectiveness Models.* Pharmacoeconomics, 2017. **35**(Suppl 1): p. 21-31.
9. Culyer, A.J., *The dictionary of health economics*. 2010: Edward Elgar Publishing.
10. Stinnett, A.A. and J. Mullahy, *Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis.* Medical decision making, 1998. **18**(2_suppl): p. S68-S80.
11. Lancsar, E. and J. Louviere, *Conducting discrete choice experiments to inform healthcare decision making: a user's guide.* Pharmacoeconomics, 2008. **26**: p. 661-677.
12. Mehrez, A. and A. Gafni, *Quality-adjusted life years, utility theory, and healthy-years equivalents.* Medical decision making, 1989. **9**(2): p. 142-149.
13. Murray, C.J. and A.D. Lopez, *Evidence-based health policy—lessons from the Global Burden of Disease Study.* Science, 1996. **274**(5288): p. 740-743.
14. WHO, G., *WHO methods and data sources for global burden of disease estimates 2000–2011.* Geneva: Department of Health Statistics and Information Systems, 2013.
15. Gold, M.R., *Cost-effectiveness in health and medicine*. 1996: Oxford university press.
16. Williams, A., *Economics of coronary artery bypass grafting.* Br Med J (Clin Res Ed), 1985. **291**(6491): p. 326-329.
17. Wright, D.R., et al., *Methods for measuring temporary health states for cost-utility analyses.* Pharmacoeconomics, 2009. **27**: p. 713-723.
18. Towers, I., A. Spencer, and J. Brazier, *Healthy year equivalents versus quality-adjusted life years: the debate continues.* Expert review of pharmacoeconomics & outcomes research, 2005. **5**(3): p. 245-254.

19. Haslett, D.W., *What is Utility?* Economics & Philosophy, 1990. **6**(1): p. 65-94.
20. Peeters, Y. and A.M. Stiggelbout, *Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities.* Value in Health, 2010. **13**(2): p. 306-309.
21. Bremner, K.E., et al., *A review and meta-analysis of prostate cancer utilities.* Medical Decision Making, 2007. **27**(3): p. 288-298.
22. Rowen, D., et al., *Comparison of general population, patient, and carer utility values for dementia health states.* Medical Decision Making, 2015. **35**(1): p. 68-80.
23. Ubel, P.A., G. Loewenstein, and C. Jepson, *Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public.* Quality of life Research, 2003. **12**: p. 599-607.
24. Rowen, D., et al., *International regulations and recommendations for utility data for health technology assessment.* Pharmacoeconomics, 2017. **35**(Suppl 1): p. 11-19.
25. Lundin, D., et al., *Guidelines of the Pharmaceutical Benefits Board for health economics evaluations. Cost-efficiency analysis from a national perspective.* Lakartidningen, 2006. **103**(47): p. 3716-3718.
26. Sanders, G.D., et al., *Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine.* Jama, 2016. **316**(10): p. 1093-1103.
27. McTaggart-Cowan, H., *Elicitation of informed general population health state utility values: a review of the literature.* Value in Health, 2011. **14**(8): p. 1153-1157.
28. Clarke, A., et al., *The effect of assessment method and respondent population on utilities elicited for Gaucher disease.* Quality of Life Research, 1997. **6**.
29. Glied, S. and P.C. Smith, *The Oxford handbook of health economics*. 2013: Oxford University Press.
30. Rowen, D., et al., *The role of condition-specific preference-based measures in health technology assessment.* Pharmacoeconomics, 2017. **35**: p. 33-41.
31. Yang, Y., L. Longworth, and J. Brazier, *An assessment of validity and responsiveness of generic measures of health-related quality of life in hearing impairment.* Qual Life Res, 2013. **22**(10): p. 2813-28.
32. Chisholm, D., A. Healey, and M. Knapp, *QALYs and mental health care.* Social psychiatry and psychiatric epidemiology, 1997. **32**: p. 68-75.
33. Ryan, M., et al., *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): p. 927-944.
34. Chalkidou, K., et al., *Cost-effective public health guidance: asking questions from the decision-maker's viewpoint.* Health economics, 2008. **17**(3): p. 441-448.
35. Al-Janabi, H., T. N Flynn, and J. Coast, *Development of a self-report measure of capability wellbeing for adults: the ICECAP-A.* Quality of life research, 2012. **21**: p. 167-176.
36. Netten, A., et al., *Outcomes of social care for adults: developing a preference-weighted measure.* Health technology assessment, 2012. **16**(16): p. 1-166.
37. Davidson, T. and L.A. Levin, *Is the societal approach wide enough to include relatives? Incorporating relatives' costs and effects in a cost-effectiveness analysis.* Appl Health Econ Health Policy, 2010. **8**(1): p. 25-35.

38. Pietrzyk, I. and M. Erdmann, *Investigating the impact of interventions on educational disparities: Estimating average treatment effects (ATEs) is not sufficient.* Research in Social Stratification and Mobility, 2020. **65**: p. 100471.
39. Peasgood, T., et al., *What is the best approach to adopt for identifying the domains for a new measure of health, social care and carer-related quality of life to measure quality-adjusted life years? Application to the development of the EQ-HWB?* Eur J Health Econ, 2021. **22**(7): p. 1067-1081.
40. Brazier, J., et al., *Well-Being: What Is It, How Does it Compare to Health and What are the Implications of Using it to Inform Health Policy?* Value in Health, 2016. **19**(7): p. A389.
41. Wisløff, T., et al., *Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010.* Pharmacoeconomics, 2014. **32**: p. 367-375.
42. Drummond, M., et al., *Methods for the economic evaluation of health care programmes*. Fourth edition / Michael F. Drummond, Mark J. Sculpher, Karl Claxton, Greg L. Stoddart, George W. Torrance. ed, ed. M.J. Sculpher, et al. 2015, Oxford, United Kingdom: Oxford, United Kingdom : Oxford University Press, 2015.
43. Shroufi, A., et al., *Measuring health: a practical challenge with a philosophical solution?* Maturitas, 2011. **68**(3): p. 210-6.
44. Bansback, N., et al., *Using a discrete choice experiment to estimate health state utility values.* J Health Econ, 2012. **31**(1): p. 306-18.
45. Ludwig, K., J.-M.G. von der Schulenburg, and W. Greiner, *Valuation of the EQ-5D-5L with composite time trade-off for the German population–an exploratory study.* Health and quality of life outcomes, 2017. **15**: p. 1-13.
46. Bell, D.E. and P.H. Farquhar, *Perspectives on utility theory.* Operations Research, 1986: p. 179-183.
47. Von Neumann, J. and O. Morgenstern, *Theory of games and economic behavior*. 60th anniversary / with an introduction by Harold W. Kuhn /and an afterword by Ariel Rubinstein. ed, ed. O. Morgenstern. 2007, Princeton, N.J. Woodstock: Princeton, N.J. Woodstock : Princeton University Press, 2007.
48. Torrance, G.W., *Measurement of health state utilities for economic appraisal: a review.* Journal of health economics, 1986. **5**(1): p. 1-30.
49. Feeny, D., et al., *Comparing directly measured standard gamble scores to HUI2 and HUI3 utility scores: group- and individual-level comparisons.* Soc Sci Med, 2004. **58**(4): p. 799-809.
50. Bali, T.G. and S. Murray, *Does Risk-Neutral Skewness Predict the Cross-Section of Equity Option Portfolio Returns?* Journal of Financial and Quantitative Analysis, 2013. **48**(4): p. 1145-1171.
51. Lugnér, A.K. and P.F.M. Krabbe, *An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health.* Expert Review of Pharmacoeconomics & Outcomes Research, 2020. **20**(4): p. 331-342.
52. Devlin, N.J. and R. Brooks, *EQ-5D and the EuroQol Group: Past, Present and Future.* Appl Health Econ Health Policy, 2017. **15**(2): p. 127-137.
53. Oppe, M., et al., *A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol.* Value Health, 2014. **17**(4): p. 445-53.

54. Janssen, B.M., et al., *Introducing the composite time trade-off: a test of feasibility and face validity.* The European Journal of Health Economics, 2013. **14**: p. 5-13.
55. Dolan, P. and C. Gudex, *Time preference, duration and health state valuations.* Health economics, 1995. **4**(4): p. 289-299.
56. Aronson, E., *The theory of cognitive dissonance: A current perspective*, in *Advances in experimental social psychology*. 1969, Elsevier. p. 1-34.
57. Shah, K., et al., *An empirical study of two alternative comparators for use in time trade-off studies.* Value in Health, 2016. **19**(1): p. 53-59.
58. Hayes, M., *Experimental development of the graphic rating method.* Psychological Bulletin, 1921. **18**: p. 98-99.
59. Flynn, D., P. Van Schaik, and A. Van Wersch, *A comparison of multi-item likert and visual analogue scales for the assessment of transactionally defined coping function1.* European Journal of Psychological Assessment, 2004. **20**(1): p. 49-58.
60. Kaplan, R.M. and J.P. Anderson, *A general health policy model: update and applications.* Health services research, 1988. **23**(2): p. 203.
61. Brazier, J. and J. Ratcliffe, *Measurement and valuation of health for economic evaluation*, in *International encyclopedia of public health*. 2017, Elsevier Inc. p. 586-593.
62. Feng, Y., D. Parkin, and N.J. Devlin, *Assessing the performance of the EQ-VAS in the NHS PROMs programme.* Quality of Life Research, 2014. **23**: p. 977-989.
63. George, B., A. Harris, and A. Mitchell, *Cost-effectiveness analysis and the consistency of decision making: evidence from pharmaceutical reimbursement in Australia (1991 to 1996).* Pharmacoeconomics, 2001. **19**: p. 1103-1109.
64. Alriksson, S. and T. Oberg, *Conjoint analysis for environmental evaluation--a review of methods and applications.* Environ Sci Pollut Res Int, 2008. **15**(3): p. 244-57.
65. Krantz, D.H. and A. Tversky, *Conjoint-measurement analysis of composition rules in psychology.* Psychological review, 1971. **78**(2): p. 151.
66. Koppen, M., *Characterization theorems in random utility theory.* Smelser, NJ; Baltes, PB (ed.), International Encyclopedia of the Social & Behavioral Sciences, Section Mathematics and Computer Sciences, 2001: p. 1646-1651.
67. Clark, M.D., et al., *Discrete choice experiments in health economics: a review of the literature.* Pharmacoeconomics, 2014. **32**(9): p. 883-902.
68. Dolan, P., et al., *Valuing health states: a comparison of methods.* Journal of health economics, 1996. **15**(2): p. 209-231.
69. Furlong, W., et al., *Guide to design and development of health-state utility instrumentation*. 1992, Centre for Health Economics and Policy Analysis (CHEPA), McMaster University ….
70. Craig, B.M., J.J. Busschbach, and J.A. Salomon, *Keep it simple: ranking health states yields values similar to cardinal measurement approaches.* Journal of clinical epidemiology, 2009. **62**(3): p. 296-305.
71. McFadden, D., *Conditional logit analysis of qualitative choice behavior.* 1973.
72. Wittenberg, E., et al., *Using Best-Worst Scaling to Understand Patient Priorities: A Case Example of Papanicolaou Tests for Homeless Women.* Ann Fam Med, 2016. **14**(4): p. 359-64.

73. Louviere, J.J. and G.G. Woodworth, *Best worst scaling: A model for largest difference judgments [Working Paper].* Faculty of Business, 1990.
74. Flynn, T.N., et al., *Best–worst scaling: what it can do for health care research and how to do it.* Journal of health economics, 2007. **26**(1): p. 171-189.
75. Coast, J. and S. Horrocks, *Developing attributes and levels for discrete choice experiments using qualitative methods.* J Health Serv Res Policy, 2007. **12**(1): p. 25-30.
76. Ratcliffe, J., et al., *Whose values in health? An empirical comparison of the application of adolescent and adult values for the CHU-9D and AQOL-6D in the Australian adolescent general population.* Value in Health, 2012. **15**(5): p. 730-736.
77. Osman, A., et al., *PNS100 Comparison of Test-Rest Reliability of Cardinal and Ordinal Preference Elicitation Methods.* Value in Health, 2021. **24**: p. S191.
78. Parkin, D. and N. Devlin, *Is there a case for using visual analogue scale valuations in cost-utility analysis?* Health economics, 2006. **15**(7): p. 653-664.
79. Louviere, J.J. and G. Woodworth, *Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data.* Journal of marketing research, 1983. **20**(4): p. 350-367.
80. Carson, R.T. and J.J. Louviere, *A Common Nomenclature for Stated Preference Elicitation Approaches.* Environmental and Resource Economics, 2011. **49**(4): p. 539-559.
81. Keeney, R.L. and H. Raiffa, *Decisions with multiple objectives: preferences and value trade-offs*. 1993: Cambridge university press.
82. Hakim, Z. and D.S. Pathak, *Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modelling.* Health Economics, 1999. **8**(2): p. 103-116.
83. Craig, B.M., D.S. Brown, and B.B. Reeve, *Valuation of child behavioral problems from the perspective of US adults.* Medical decision making, 2016. **36**(2): p. 199-209.
84. Mulhern, B., et al., *One Method, Many Methodological Choices: A Structured Review of Discrete-Choice Experiments for Health State Valuation.* Pharmacoeconomics, 2019. **37**(1): p. 29-43.
85. Street, D.J. and L. Burgess, *The construction of optimal stated choice experiments: theory and methods*. 2007: John Wiley & Sons.
86. Jonker, M.F., et al., *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): p. 1534-1547.
87. Rowen, D., J. Brazier, and B. Van Hout, *A comparison of methods for converting DCE values onto the full health-dead QALY scale.* Med Decis Making, 2015. **35**(3): p. 328-40.
88. Hensher, D.A., J.M. Rose, and W.H. Greene, *Applied choice analysis: a primer*. 2005: Cambridge university press.
89. Louviere, J.J., et al., *Designing Discrete Choice Experiments: Do Optimal Designs Come at a Price?* Journal of Consumer Research, 2008. **35**(2): p. 360-375.
90. DeShazo, J. and G. Fermo, *Designing choice sets for stated preference methods: the effects of complexity on choice consistency.* Journal of Environmental Economics and management, 2002. **44**(1): p. 123-143.
91. Kuhfeld, W.F., *Experimental design, efficiency, coding, and choice designs.* Marketing research methods in sas: Experimental design, choice, conjoint, and graphical techniques, 2005: p. 47-97.

92. Burgess, L., D.J. Street, and N. Wasi, *Comparing designs for choice experiments: a case study.* Journal of Statistical Theory and Practice, 2011. **5**(1): p. 25-46.
93. Emery, D.R. and F. Hutton Barron, *Axiomatic and numerical conjoint measurement: an evaluation of diagnostic efficacy.* Psychometrika, 1979. **44**: p. 195-210.
94. Ryan, M., et al., *Use of discrete choice experiments to elicit preferences.* BMJ Quality & Safety, 2001. **10**(suppl 1): p. i55-i60.
95. Eftekhari, H., M. Banerjee, and Y.a. Ritov, *Design of c-optimal experiments for high-dimensional linear models.* Bernoulli, 2023. **29**(1): p. 652-668.
96. Harman, R. and T. Jurík, *Computing c-optimal experimental designs using the simplex method of linear programming.* Computational statistics & data analysis, 2008. **53**(2): p. 247-254.
97. McIntosh, E., *Applied methods of cost-benefit analysis in health care [electronic resource]*, ed. E. McIntosh. 2010, Oxford: Oxford University Press, 2010.
98. Carlsson, F. and P. Martinsson, *Design techniques for stated preference methods in health economics.* Health economics, 2003. **12**(4): p. 281-294.
99. Street, D.J., L. Burgess, and J.J. Louviere, *Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments.* International Journal of Research in Marketing, 2005. **22**(4): p. 459-470.
100. Hanson, W.E., et al., *Mixed methods research designs in counseling psychology.* Journal of counseling psychology, 2005. **52**(2): p. 224.
101. Mangham, L.J. and K. Hanson, *Employment preferences of public sector nurses in Malawi: results from a discrete choice experiment.* Tropical Medicine & International Health, 2008. **13**(12): p. 1433-1441.
102. Janssen, E.M., J.B. Segal, and J.F. Bridges, *A framework for instrument development of a choice experiment: an application to type 2 diabetes.* The Patient-Patient-Centered Outcomes Research, 2016. **9**: p. 465-479.
103. Thai, T.T.H., et al., *A Comparison of Full and Partial Choice Set Designs in a Labelled Discrete Choice Experiment.* The Patient, 2021. **14**(6): p. 866-867.
104. de Bekker-Grob, E.W., et al., *Sample size requirements for discrete-choice experiments in healthcare: a practical guide.* The Patient-Patient-Centered Outcomes Research, 2015. **8**: p. 373-384.
105. Dillman, D.A., *The promise and challenge of pushing respondents to the web in mixed-mode surveys.* Survey Methodology, 2017. **43**(1): p. 3-31.
106. Lancaster, K.J., *A new approach to consumer theory.* Journal of political economy, 1966. **74**(2): p. 132-157.
107. Ali, S. and S. Ronaldson, *Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods.* British medical bulletin, 2012. **103**(1): p. 21-44.
108. Von Winterfeldt, D. and G.W. Fischer, *Multi-Attribute Utility Theory: Models and Assessment Procedures*, in *Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973*, D. Wendt and C. Vlek, Editors. 1975, Springer Netherlands: Dordrecht. p. 47-85.

109. Schneider, P., et al., *The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states [version 1; peer review: 2 approved, 1 approved with reservations].* Wellcome Open Research, 2022. **7**(14).
110. Craig, B.M., J.J. Busschbach, and J.A. Salomon, *Modeling ranking, time trade-off and visual analogue scale values for EQ-5D health states: A review and comparison of methods.* Medical care, 2009. **47**(6): p. 634.
111. Golman, R., *Homogeneity bias in models of discrete choice with bounded rationality.* Journal of Economic Behavior & Organization, 2012. **82**(1): p. 1-11.
112. Goossens, L., et al., *Inpatient hospital care or hospital-at-home for COPD exacerbations: A discrete choice experiment*. 2012, Eur Respiratory Soc.
113. Fosgerau, M. and S. Hess, *A comparison of methods for representing random taste heterogeneity in discrete choice models.* European Transport-Trasporti Europei, 2009. **42**: p. 1-25.
114. Hanmer, J., et al., *Evaluation of options for presenting health-states from PROMIS® item banks for valuation exercises.* Quality of life research, 2018. **27**: p. 1835-1843.
115. Kjaer, T. and D. Gyrd-Hansen, *Preference heterogeneity and choice of cardiac rehabilitation program: results from a discrete choice experiment.* Health policy, 2008. **85**(1): p. 124-132.
116. de Bekker-Grob, E.W., M. Ryan, and K. Gerard, *Discrete choice experiments in health economics: a review of the literature.* Health economics, 2012. **21**(2): p. 145-172.
117. Chapaaan, R.G. and R. Staelin, *Exploiting rank ordered choice set data within the stochastic utility model.* Journal of marketing research, 1982. **19**(3): p. 288-301.
118. Hoogendoorn, M., et al., *Exploring the Impact of Adding a Respiratory Dimension to the EQ-5D-5L.* Med Decis Making, 2019. **39**(4): p. 393-404.
119. Stolk, E., et al., *Overview, Update, and Lessons Learned From the International EQ-5D-5L Valuation Work: Version 2 of the EQ-5D-5L Valuation Protocol.* Value Health, 2019. **22**(1): p. 23-30.
120. Salomon, J.A., *Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data.* Population health metrics, 2003. **1**(1): p. 1-12.
121. McCabe, C., et al., *Using rank data to estimate health state utility models.* Journal of health economics, 2006. **25**(3): p. 418-431.
122. Brazier, J.E., et al., *Developing a New Version of the SF-6D Health State Classification System From the SF-36v2: SF-6Dv2.* Med Care, 2020. **58**(6): p. 557-565.
123. Wang, H., et al., *Discrete Choice Experiments in Health State Valuation: A Systematic Review of Progress and New Trends*. Applied Health Economics and Health Policy, 2023: p. 1-14.
124. Bansback, N., et al., *Testing a discrete choice experiment including duration to value health states for large descriptive systems: Addressing design and sampling issues.* Social Science & Medicine, 2014. **114**: p. 38-48.
125. Himmler, S., et al., *Estimating an anchored utility tariff for the well-being of older people measure (WOOP) for the Netherlands.* Soc Sci Med, 2022. **301**: p. 114901.
126. Lim, S., et al., *Severity-Stratified Discrete Choice Experiment Designs for Health State Evaluations.* Pharmacoeconomics, 2018. **36**(11): p. 1377-1389.

127. Flynn, T.N., et al., *Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale.* Population Health Metrics, 2008. **6**(1): p. 6.
128. Luce, R.D., *Individual choice behavior: A theoretical analysis*. 2012: Courier Corporation.
129. Norman, R., B. Mulhern, and R. Viney, *The Impact of Different DCE-Based Approaches When Anchoring Utility Scores.* Pharmacoeconomics, 2016. **34**(8): p. 805-14.
130. Bleichrodt, H., *A new explanation for the difference between time trade-off utilities and standard gamble utilities.* Health economics, 2002. **11**(5): p. 447-456.
131. Anderl, C., et al., *Cooperative preferences fluctuate across the menstrual cycle.* Judgment and Decision Making, 2015. **10**(5): p. 400-406.
132. McCabe, A. and C. Peterson, *Developing narrative structure*. 1991: Psychology Press.
133. Salomon, J.A., et al., *Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010.* The Lancet, 2012. **380**(9859): p. 2129-2143.
134. Bijlenga, D., E. Birnie, and G.J. Bonsel, *Feasibility, reliability, and validity of three health-state valuation methods using multiple-outcome vignettes on moderate-risk pregnancy at term.* Value in Health, 2009. **12**(5): p. 821-827.
135. Comans, T.A., et al., *Valuing the AD-5D Dementia Utility Instrument: An Estimation of a General Population Tariff.* Pharmacoeconomics, 2020. **38**(8): p. 871-881.
136. Craig, B.M., et al., *US valuation of health outcomes measured using the PROMIS-29.* Value Health, 2014. **17**(8): p. 846-53.
137. Sullivan, P.W., W.F. Lawrence, and V. Ghushchyan, *A national catalog of preference-based scores for chronic conditions in the United States.* Medical care, 2005: p. 736-749.
138. Wang, H., D.A. Kindig, and J. Mullahy, *Variation in Chinese population health related quality of life: results from a EuroQol study in Beijing, China.* Quality of life research, 2005. **14**: p. 119-132.
139. Jyani, G., et al., *Development of an EQ-5D Value Set for India Using an Extended Design (DEVINE) Study: The Indian 5-Level Version EQ-5D Value Set.* Value Health, 2022. **25**(7): p. 1218-1226.
140. Fox-Rushby, J., *First steps to assessing semantic equivalence of the EuroQol instrument: Results of a questionnaire survey to members of the EuroQol Group*, in *EQ-5D concepts and methods: A developmental history*. 2005, Springer. p. 35-52.
141. Longworth, L., J. Singh, and J. Brazier, *An evaluation of the performance of EQ-5D: a review of reviews of psychometric properties.* Value in Health, 2014. **17**(7): p. A570.
142. Mulhern, B., et al., *Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets.* Pharmacoeconomics, 2018. **36**: p. 699-713.
143. Wille, N., et al., *Development of the EQ-5D-Y: a child-friendly version of the EQ-5D.* Quality of life research, 2010. **19**(6): p. 875-886.
144. Brazier, J. and A. Tsuchiya, *Improving Cross-Sector Comparisons: Going Beyond the Health-Related QALY.* Appl Health Econ Health Policy, 2015. **13**(6): p. 557-65.
145. Williams, A., *The measurement and valuation of health: a chronicle*. 1995: University of York York.
146. Dolan, P., et al., *A social tariff for EuroQol: results from a UK general population survey*. 1995.

147. Oppe, M., N.J. Devlin, and A. Szende, *EQ-5D value sets: inventory, comparative review and user guide*. 2007: Springer.
148. Devlin, N.J., et al., *Valuing health-related quality of life: An EQ-5D-5L value set for England.* Health Econ, 2018. **27**(1): p. 7-22.
149. Lipman, S.A., et al., *Time and lexicographic preferences in the valuation of EQ-5D-Y with time trade-off methodology.* The European Journal of Health Economics, 2023. **24**(2): p. 293-305.
150. Scott, A., *Identifying and analysing dominant preferences in discrete choice experiments: an application in health care.* Journal of economic Psychology, 2002. **23**(3): p. 383-398.
151. Viney, R., E. Lancsar, and J. Louviere, *Discrete choice experiments to measure consumer preferences for health and healthcare.* Expert review of pharmacoeconomics & outcomes research, 2002. **2**(4): p. 319-326.
152. Ramos-Goñi, J.M., et al., *Quality control process for EQ-5D-5L valuation studies.* Value in health, 2017. **20**(3): p. 466-473.
153. Brooks, R., K.S. Boye, and B. Slaap, *EQ-5D: a plea for accurate nomenclature.* Journal of Patient-Reported Outcomes, 2020. **4**(1): p. 1-3.
154. Cole, A., et al., *Valuing EQ-5D-5L health states 'in context' using a discrete choice experiment.* Eur J Health Econ, 2018. **19**(4): p. 595-605.
155. van Hout, B., et al., *Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets.* Value Health, 2012. **15**(5): p. 708-15.
156. Lamers, L., et al., *Comparison of EQ-5D and SF-6D utilities in mental health patients.* Health economics, 2006. **15**(11): p. 1229-1236.
157. Saarni, S.I., et al., *Quality of life of people with schizophrenia, bipolar disorder and other psychotic disorders.* The British Journal of Psychiatry, 2010. **197**(5): p. 386-394.
158. Finch, A.P., J.E. Brazier, and C. Mukuria, *Selecting bolt-on dimensions for the EQ-5D: examining their contribution to health-related quality of life.* Value in Health, 2019. **22**(1): p. 50-61.
159. Brazier, J.E., et al., *Future directions in valuing benefits for estimating QALYs: is time up for the EQ-5D?* Value in Health, 2019. **22**(1): p. 62-68.
160. Norman, R. and J.A. Olsen, *Expanding the scope of value for economic evaluation: the EQ-HWB.* Value in Health, 2022. **25**(4): p. 480-481.
161. Brazier, J., et al., *The EQ-HWB: overview of the development of a measure of health and wellbeing and key results.* Value in Health, 2022. **25**(4): p. 482-491.
162. Mukuria, C., et al., *Qualitative Review on Domains of Quality of Life Important for Patients, Social Care Users, and Informal Carers to Inform the Development of the EQ-HWB.* Value Health, 2022. **25**(4): p. 492-511.
163. Brazier, J., T. Peasgood, and C. Mukuria, *Development of a new generic measure of health and wellbeing for estimating Quality Adjusted Life Years: the EQ Health wellbeing (EQ-HWB).* Value Health.
164. Peasgood, T., et al., *Criteria for item selection for a preference-based measure for use in economic evaluation.* Qual Life Res, 2021. **30**(5): p. 1425-1432.
165. Carlton, J., et al., *Generation, Selection, and Face Validation of Items for a New Generic Measure of Quality of Life: The EQ-HWB.* Value Health, 2022. **25**(4): p. 512-524.

166. Peasgood, T., et al., *Developing a New Generic Health and Wellbeing Measure: Psychometric Survey Results for the EQ-HWB.* Value Health, 2022. **25**(4): p. 525-533.
167. Mukuria, C., et al., *Valuing the EQ Health and Wellbeing Short Using Time Trade-Off and a Discrete Choice Experiment: A Feasibility Study.* Value in Health, 2023.
168. Mukuria, C., T. Peasgood, and J. Brazier, *Applying EuroQol Portable Valuation Technology to the EQ Health and Wellbeing Short (EQHWB-S): A pilot study.* School of Health and Related Research, University of Sheffield Discussion Paper Series, 2021.
169. Lee, P., et al., *Exploring the Comparability Between EQ-5D and the EQ Health and Wellbeing in the General Australian Population.* Value in Health, 2024.
170. Masutti, S., et al., *Content validity of the EQ-HWB and EQ-HWB-S in a sample of Italian patients, informal caregivers and members of the general public.* Journal of Patient-Reported Outcomes, 2024. **8**(1): p. 36.
171. Bailey, C., et al., *The Validity of the EuroQol Health and Wellbeing Short Version (EQ-HWB-S) Instrument in Parents of Children With and Without Health Conditions.* PharmacoEconomics, 2024: p. 1-17.
172. Peasgood, T., et al., *A conceptual comparison of well-being measures used in the UK.* 2014.
173. Peasgood, T., J. Carlton, and J. Brazier, *A Qualitative Study of the Views of Health and Social Care Decision-Makers on the Role of Wellbeing in Resource Allocation Decisions in the UK.* Economies, 2019. **7**(1): p. 14.
174. Soekhai, V., et al., *Discrete Choice Experiments in Health Economics: Past, Present and Future.* Pharmacoeconomics, 2019. **37**(2): p. 201-226.
175. Bahrampour, M., et al., *Discrete choice experiments to generate utility values for multi-attribute utility instruments: a systematic review of methods.* Eur J Health Econ, 2020. **21**(7): p. 983-992.
176. Prosser, L.A., et al., *Using a discrete choice experiment to elicit time trade-off and willingness-to-pay amounts for influenza health-related quality of life at different ages.* Pharmacoeconomics, 2013. **31**: p. 305-315.
177. Wu, J., et al., *Valuation of SF-6Dv2 Health States in China Using Time Trade-off and Discrete-Choice Experiment with a Duration Dimension.* Pharmacoeconomics, 2021.
178. Bridges, J.F., et al., *Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force.* Value Health, 2011. **14**(4): p. 403-13.
179. Louviere, J.J., T.N. Flynn, and R.T. Carson, *Discrete Choice Experiments Are Not Conjoint Analysis*. Journal of Choice Modelling, 2010. **3**(3): p. 57-72.
180. Flynn, T.N., *Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling.* Expert review of pharmacoeconomics & outcomes research, 2010. **10**(3): p. 259-267.
181. *The BWS multi-profile case*, in *Best-Worst Scaling: Theory, Methods and Applications*, A.A.J. Marley, J.J. Louviere, and T.N. Flynn, Editors. 2015, Cambridge University Press: Cambridge. p. 89-113.
182. Mott, D.J., et al., *Valuing EQ-5D-Y-3L Health States Using a Discrete Choice Experiment: Do Adult and Adolescent Preferences Differ?* Med Decis Making, 2021: p. 272989X21999607.

183. Krabbe, P., et al., *A two-step procedure to generate utilities for the Infant health-related Quality of life Instrument (IQI).* PLoS One, 2020. **15**(4): p. e0230852.
184. Fenwick, E.K., et al., *Validation of a novel diabetic retinopathy utility index using discrete choice experiments.* Br J Ophthalmol, 2020. **104**(2): p. 188-193.
185. Ratcliffe, J., et al., *Valuing the Quality-of-Life Aged Care Consumers (QOL-ACC) Instrument for Quality Assessment and Economic Evaluation.* Pharmacoeconomics, 2022. **40**(11): p. 1069-1079.
186. O'Hara, J., et al., *Evidence of a disability paradox in patient-reported outcomes in haemophilia.* Haemophilia, 2021.
187. King, M.T., et al., *QLU-C10D: a health state classification system for a multi-attribute utility measure based on the EORTC QLQ-C30.* Quality of Life Research, 2016. **25**(3): p. 625-636.
188. Dams, J., et al., *German tariffs for the ICECAP-Supportive Care Measure (ICECAP-SCM) for use in economic evaluations at the end of life.* Eur J Health Econ, 2021. **22**(3): p. 365-380.
189. Baji, P., et al., *Development of Population Tariffs for the CarerQol Instrument for Hungary, Poland and Slovenia: A Discrete Choice Experiment Study to Measure the Burden of Informal Caregiving.* Pharmacoeconomics, 2020. **38**(6): p. 633-643.
190. Voormolen, D.C., et al., *Health-related quality of life after traumatic brain injury: deriving value sets for the QOLIBRI-OS for Italy, The Netherlands and The United Kingdom.* Qual Life Res, 2020. **29**(11): p. 3095-3107.
191. Bahrampour, M., et al., *Utility Values for the CP-6D, a Cerebral Palsy-Specific Multi-Attribute Utility Instrument, Using a Discrete Choice Experiment.* Patient, 2021. **14**(1): p. 129-138.
192. Bailey, C., et al., *'The ICECAP-SCM tells you more about what I'm going through': A think-aloud study measuring quality of life among patients receiving supportive and palliative care.* Palliative Medicine, 2016. **30**(7): p. 642-652.
193. King, M.T., et al., *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): p. 225-238.
194. Norman, R., et al., *U.K. utility weights for the EORTC QLU-C10D.* Health Econ, 2019. **28**(12): p. 1385-1401.
195. Shah, K.K., et al., *An exploration of methods for obtaining 0 = dead anchors for latent scale EQ-5D-Y values.* Eur J Health Econ, 2020. **21**(7): p. 1091-1103.
196. Mulhern, B., et al., *Using Discrete Choice Experiments with Duration to Model EQ-5D-5L Health State Preferences.* Med Decis Making, 2017. **37**(3): p. 285-297.
197. Rowen, D., et al., *Estimating a Preference-Based Single Index Measuring the Quality-of-Life Impact of Self-Management for Diabetes.* Med Decis Making, 2018. **38**(6): p. 699-707.
198. Shafie, A.A., et al., *EQ-5D-5L Valuation for the Malaysian Population.* Pharmacoeconomics, 2019. **37**(5): p. 715-725.
199. Hauber, A.B., et al., *Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force.* Value in Health, 2016. **19**(4): p. 300-315.

200. Manahan, R., et al., *PSAT097 Patient Preference Research: Preferred Adjunctive Medication Attributes of Adult Patients with Classic Congenital Adrenal Hyperplasia.* Journal of the Endocrine Society, 2022. **6**(Supplement_1): p. A118-A118.
201. Augustovski, F., et al., *Peruvian Valuation of the EQ-5D-5L: A Direct Comparison of Time Trade-Off and Discrete Choice Experiments.* Value Health, 2020. **23**(7): p. 880-888.
202. Mulhern, B., et al., *Investigating the relative value of health and social care related quality of life using a discrete choice experiment.* Social science & medicine, 2019. **233**: p. 28-37.
203. Norman, R., et al., *Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format.* Qual Life Res, 2016. **25**(3): p. 637-49.
204. Mulhern, B.J., et al., *Valuing the SF-6Dv2 Classification System in the United Kingdom Using a Discrete-choice Experiment With Duration.* Med Care, 2020. **58**(6): p. 566-573.
205. Reed Johnson, F., et al., *Constructing experimental designs for discrete-choice experiments: report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force.* Value Health, 2013. **16**(1): p. 3-13.
206. Walker, J.L., et al., *D-efficient or deficient? A robustness analysis of stated choice experimental designs.* Theory and Decision, 2018. **84**(2): p. 215-238.
207. Shiroiwa, T., et al., *Valuation Survey of EQ-5D-Y Based on the International Common Protocol: Development of a Value Set in Japan.* Med Decis Making, 2021. **41**(5): p. 597-606.
208. Mulhern, B., et al., *Preparatory study for the revaluation of the EQ-5D tariff: methodology report.* Health Technol Assess, 2014. **18**(12): p. vii-xxvi, 1-191.
209. Jonker, M.F., et al., *Effect of level overlap and color coding on attribute non-attendance in discrete choice experiments.* Value in Health, 2018. **21**(7): p. 767-771.
210. Vanniyasingam, T., et al., *Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments.* BMJ open, 2016. **6**(7): p. e011985.
211. Louviere, J.J., *Hierarchical information integration: A new method for the design and analysis of complex multlattribute judgment problems.* ACR North American Advances, 1984.
212. Witt, J., A. Scott, and R.H. Osborne, *Designing choice experiments with many attributes. An application to setting priorities for orthopaedic waiting lists.* Health Econ, 2009. **18**(6): p. 681-96.
213. Chrzan, K., *Using partial profile choice experiments to handle large numbers of attributes.* International Journal of Market Research, 2010. **52**(6): p. 827-840.
214. Nielsen, H. and P. Amer, *An approach to derive economic weights in breeding objectives using partial profile choice experiments.* Animal, 2007. **1**(9): p. 1254-1262.
215. Hansen, P. and F. Ombler, *A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives.* Journal of Multi-Criteria Decision Analysis, 2008. **15**(3-4): p. 87-107.
216. Punkka, A. and A. Salo, *Preference programming with incomplete ordinal information.* European Journal of Operational Research, 2013. **231**(1): p. 141-150.
217. Salo, A. and A. Punkka, *Rank inclusion in criteria hierarchies.* European Journal of Operational Research, 2005. **163**(2): p. 338-356.

218. Larichev, O. and H. Moshkovich, *ZAPROS-LM—a method and system for ordering multiattribute alternatives*. European Journal of Operational Research, 1995. **82**(3): p. 503-521.
219. Sullivan, T., et al., *A new tool for creating personal and social EQ-5D-5L value sets, including valuing 'dead'.* Soc Sci Med, 2020. **246**: p. 112707.
220. Hoefman, R.J., et al., *A Discrete Choice Experiment to Obtain a Tariff for Valuing Informal Care Situations Measured with the CarerQol Instrument.* Medical Decision Making, 2014. **34**(1): p. 84-96.
221. Jonker, M.F., et al., *Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments.* Health Econ, 2019. **28**(3): p. 350-363.
222. dos Santos Navarro, R.d.C., et al., *Balanced incomplete block design: an alternative for data collection in the optimized descriptive profile.* Food research international, 2014. **64**: p. 289-297.
223. Jaynes, J., W.-K. Wong, and H. Xu, *Using blocked fractional factorial designs to construct discrete choice experiments for healthcare studies*. Statistics in Medicine, 2016. **35**(15): p. 2543-2560.
224. Kessels, R., B. Jones, and P. Goos, *A comparison of partial profile designs for discrete choice experiments with an application in software development.* 2012.
225. Malik, M., et al., *A Pilot Study of Valuation Methods of the EQ-5D and the Impact of Literacy, Cultural and Religious Factors on Preferences*. Value Health Reg Issues, 2022. **30**: p. 48-58.
226. Rowen, D., et al., *Assessing the comparative feasibility, acceptability and equivalence of videoconference interviews and face-to-face interviews using the time trade-off technique.* Social Science & Medicine, 2022. **309**: p. 115227.
227. de Bekker-Grob, E.W., et al., *Are healthcare choices predictable? The impact of discrete choice experiment designs and models*. Value in Health, 2019. **22**(9): p. 1050-1062.
228. Doherty, E., et al., *An Exploration on Attribute Non-attendance Using Discrete Choice Experiment Data from the Irish EQ-5D-5L National Valuation Study.* Pharmacoecon Open, 2021. **5**(2): p. 237-244.
229. Wang, K., et al., *Using Eye-Tracking Technology with Older People in Memory Clinics to Investigate the Impact of Mild Cognitive Impairment on Choices for EQ-5D-5L Health States Preferences.* Appl Health Econ Health Policy, 2021. **19**(1): p. 111-121.
230. Jonker, M.F., et al., *Advocating a Paradigm Shift in Health-State Valuations: The Estimation of Time-Preference Corrected QALY Tariffs.* Value Health, 2018. **21**(8): p. 993-1001.
231. Stolk, E.A., et al., *Discrete choice modeling for the quantification of health states: the case of the EQ-5D.* Value Health, 2010. **13**(8): p. 1005-13.
232. Gotwalt, C.M., B.A. Jones, and D.M. Steinberg, *Fast Computation of Designs Robust to Parameter Uncertainty for Nonlinear Settings*. Technometrics, 2009. **51**(1): p. 88-95.
233. Kessels, R., et al., *Rejoinder: the usefulness of Bayesian optimal designs for discrete choice experiments.* Applied stochastic models in business and industry, 2011. **27**(3): p. 197-203.
234. Norman, R., et al., *Issues in the design of discrete choice experiments.* The Patient-Patient-Centered Outcomes Research, 2019. **12**: p. 281-285.

235. Karimi, M., J. Brazier, and S. Paisley, *How do individuals value health states? A qualitative investigation.* Social Science & Medicine, 2017. **172**: p. 80-88.
236. Jonker, M.F. and B. Donkers, *Interaction Effects in Health State Valuation Studies: An Optimal Scaling Approach.* Value in Health, 2023. **26**(4): p. 554-566.
237. Mulhern, B., et al., *How Should Discrete Choice Experiments with Duration Choice Sets Be Presented for the Valuation of Health States?* Med Decis Making, 2018. **38**(3): p. 306-318.
238. Ben-Akiva, M., et al., *Hybrid choice models: Progress and challenges.* Marketing Letters, 2002. **13**: p. 163-175.
239. Zhang, J., et al., *Too many attributes: A test of the validity of combining discrete-choice and best–worst scaling data.* Journal of choice modelling, 2015. **15**: p. 1-13.
240. Aguiar, M., et al., *Designing discrete choice experiments using a patient-oriented approach.* The Patient-Patient-Centered Outcomes Research, 2021. **14**(4): p. 389-397.
241. Mansfield, C., J. Sutphin, and M. Boeri, *Assessing the impact of excluded attributes on choice in a discrete choice experiment using a follow-up question.* Health Econ, 2020. **29**(10): p. 1307-1315.
242. Kangwanrattanakul, K. and W. Phimarn, *A systematic review of the development and testing of additional dimensions for the EQ-5D descriptive system.* Expert Review of Pharmacoeconomics & Outcomes Research, 2019. **19**(4): p. 431-443.
243. Yang, Y., et al., *An exploratory study to test the impact on three "bolt-on" items to the EQ-5D.* Value in Health, 2015. **18**(1): p. 52-60.
244. Louviere, J.J., D.A. Hensher, and J.D. Swait, *Stated choice methods: analysis and applications*. 2000: Cambridge university press.
245. Feng, Y.S., et al., *Psychometric properties of the EQ-5D-5L: a systematic review of the literature.* Qual Life Res, 2021. **30**(3): p. 647-673.
246. Sheffield, U.o., *Extending the QALY Project summary*. 2020: the United Kingdom.
247. Bjorner, J.B., M. Kosinski, and J.E. Ware Jr, *Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT™).* Quality of Life Research, 2003. **12**(8): p. 913-933.
248. Ray, J.V., et al., *Positive and negative item wording and its influence on the assessment of callous-unemotional traits.* Psychol Assess, 2016. **28**(4): p. 394-404.
249. Floyd, F.J. and K.F. Widaman, *Factor analysis in the development and refinement of clinical assessment instruments.* Psychological assessment, 1995. **7**(3): p. 286.
250. Mueller, R.O. and G.R. Hancock, *Factor Analysis and Latent Structure, Confirmatory*, in *International Encyclopedia of the Social & Behavioral Sciences*, N.J. Smelser and P.B. Baltes, Editors. 2001, Pergamon: Oxford. p. 5239-5244.
251. Finch, A.P., J. Brazier, and C. Mukuria, *Selecting bolt-on dimensions for the EQ-5D: testing the impact of hearing, sleep, cognition, energy, and relationships on preferences using pairwise choices.* 2021.
252. Dufresne, É., et al., *SF-6Dv2 preference value set for health utility in food allergy.* Allergy, 2021. **76**(1): p. 326-338.
253. Coast, J., et al., *Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations.* Health Econ, 2012. **21**(6): p. 730-41.
254. Watson, V., F. Becker, and E. de Bekker-Grob, *Discrete choice experiment response rates: A meta-analysis.* Health economics, 2017. **26**(6): p. 810-817.

255. Hensher, D.A., *How do respondents process stated choice experiments? Attribute consideration under varying information load.* Journal of applied econometrics, 2006. **21**(6): p. 861-878.
256. Vass, C., D. Rigby, and K. Payne, *The Role of Qualitative Research Methods in Discrete Choice Experiments.* Med Decis Making, 2017. **37**(3): p. 298-313.
257. Wong, L.P., *Focus group discussion: a tool for health and medical research.* Singapore Med J, 2008. **49**(3): p. 256-60.
258. Barbour, R.S., *The SAGE Handbook of Qualitative Data Analysis*. 2014, SAGE Publications Ltd: London.
259. Krueger, R.A., *Focus groups: A practical guide for applied research*. 2014: Sage publications.
260. Kitzinger, J., *Qualitative research: introducing focus groups.* Bmj, 1995. **311**(7000): p. 299-302.
261. Clarke, A., *Focus group interviews in health-care research.* Professional Nurse (London, England), 1999. **14**(6): p. 395-397.
262. Morse, J.M., *The significance of saturation*. 1995, Sage Publications Sage CA: Thousand Oaks, CA. p. 147-149.
263. Goossens, L.M., et al., *The fold-in, fold-out design for DCE choice tasks: application to burden of disease.* Medical Decision Making, 2019. **39**(4): p. 450-460.
264. Bazeley, P., *Qualitative data analysis: Practical strategies*. Qualitative Data Analysis, 2020: p. 1-584.
265. Smithson, J., *Using and analysing focus groups: limitations and possibilities.* International journal of social research methodology, 2000. **3**(2): p. 103-119.
266. Simon, H.A., *A behavioral model of rational choice.* The quarterly journal of economics, 1955: p. 99-118.
267. Lagarde, M., *Investigating attribute non-attendance and its consequences in choice experiments with latent class models.* Health economics, 2013. **22**(5): p. 554-567.
268. Swait, J., *A non-compensatory choice model incorporating attribute cutoffs.* Transportation Research Part B: Methodological, 2001. **35**(10): p. 903-928.
269. Swait, J., M. Popa, and L. Wang, *Capturing context-sensitive information usage in choice models via mixtures of information archetypes.* Journal of Marketing Research, 2016. **53**(5): p. 646-664.
270. Veldwijk, J., et al., *Taking the Shortcut: Simplifying Heuristics in Discrete Choice Experiments.* The Patient-Patient-Centered Outcomes Research, 2023: p. 1-15.
271. Schneider, P., *The QALY is ableist: on the unethical implications of health states worse than dead.* Quality of Life Research, 2022. **31**(5): p. 1545-1552.
272. Bergen, N. and R. Labonté, *"Everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research.* Qualitative health research, 2020. **30**(5): p. 783-792.
273. Mulhern, B., et al., *Is dimension order important when valuing health states using discrete choice experiments including duration?* Pharmacoeconomics, 2017. **35**: p. 439-451.
274. Tsuchiya, A., et al., *Using DCE with duration to examine the robustness of preferences across the five dimensions of the EuroQol instrument: The second paper from the FEDEV project*. EuroQol Group Plenary Proceedings, 2014.

275. Norman, R., et al., *Order of presentation of dimensions does not systematically bias utility weights from a discrete choice experiment.* Value in Health, 2016. **19**(8): p. 1033-1038.
276. Kjaer, T., et al., *Ordering effect and price sensitivity in discrete choice experiments: need we worry?* Health economics, 2006. **15**(11): p. 1217-1228.
277. Brazier, J., et al., *Using rank and discrete choice data to estimate health state utility values on the QALY scale.* 2009.
278. Craig, B.M. and J.J. Busschbach, *The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation.* Population Health Metrics, 2009. **7**: p. 1-10.
279. Devlin, N., B. Roudijk, and K. Ludwig, *Value sets for EQ-5D-5L: a compendium, comparative review & user guide.* 2022.
280. Marten, O., et al., *Implausible States: Prevalence of EQ-5D-5L States in the General Population and Its Effect on Health State Valuation.* Med Decis Making, 2020. **40**(6): p. 735-745.
281. van Cranenburgh, S. and A.T. Collins, *New software tools for creating stated choice experimental designs efficient for regret minimisation and utility maximisation decision rules.* Journal of Choice Modelling, 2019. **31**: p. 104-123.
282. Grover, R. and M. Vriens, *The Handbook of Marketing Research : Uses, Misuses, and Future Advances*. 2006, Thousand Oaks, UNITED STATES: SAGE Publications, Incorporated.
283. Norman, R., et al., *The Use of a Discrete Choice Experiment Including Both Duration and Dead for the Development of an EQ-5D-5L Value Set for Australia.* PharmacoEconomics, 2023. **41**(4): p. 427-438.
284. Cook, R.D. and C.J. Nachtrheim, *A comparison of algorithms for constructing exact D-optimal designs.* Technometrics, 1980. **22**(3): p. 315-324.
285. Johnson, F.R., et al., *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): p. 3-13.
286. Levy, P.S. and S. Lemeshow, *Sampling of populations: methods and applications*. 2013: John Wiley & Sons.
287. Rosseel, Y., *lavaan: An R package for structural equation modeling.* Journal of statistical software, 2012. **48**: p. 1-36.
288. Burton, A., et al., *The design of simulation studies in medical statistics.* Statistics in medicine, 2006. **25**(24): p. 4279-4292.
289. Akaike, H., *Information theory and an extension of the maximum likelihood principle*, in *Selected papers of hirotugu akaike*. 1998, Springer. p. 199-213.
290. Sawa, T., *Information criteria for discriminating among alternative regression models.* Econometrica: Journal of the Econometric Society, 1978: p. 1273-1291.
291. Ward, E.J., *A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools.* Ecological Modelling, 2008. **211**(1-2): p. 1-10.
292. Long, J.S. and J. Freese, *Regression models for categorical dependent variables using Stata*. Vol. 7. 2006: Stata press.
293. van der Pol, M. and J. Cairns, *Estimating time preferences for health using discrete choice experiments.* Social Science & Medicine, 2001. **52**(9): p. 1459-1470.

294. Veldwijk, J., et al., *The effect of including an opt-out option in discrete choice experiments.* PloS one, 2014. **9**(11): p. e111805.
295. Monzani, D., et al., *Patient Preferences for Lung Cancer Treatments: A Study Protocol for a Preference Survey Using Discrete Choice Experiment and Swing Weighting*. Front Med (Lausanne), 2021. **8**: p. 689114.
296. Ramos-Goñi, J.M., et al., *Does Changing the Age of a Child to be Considered in 3-Level Version of EQ-5D-Y Discrete Choice Experiment–Based Valuation Studies Affect Health Preferences?* Value in Health, 2022. **25**(7): p. 1196-1204.
297. Brazier, J., et al., *The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results.* Value Health, 2022. **25**(4): p. 482-491.
298. Statistics, U.O.f.N. *Estimates of the population for the UK, England, Wales, Scotland, and Northern Ireland*. 2024 26 March 2024 [cited 2024 3rd]; April]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.
299. (AIHW), A.I.o.H.a.W. *The health of Australia's females*. 2023 27 Jun 2023 [cited 2024 3rd April]; Available from: https://www.aihw.gov.au/reports/men-women/female-health/contents/who-are.
300. Welfare, A.I.o.H.a. *Population: Australian population - age and sex*. 2023 Sept. 2023 [cited 2024 3rd April]; Available from: https://www.housingdata.gov.au/visualisation/population/australian-population-age-and-sex.
301. Statistics, O.f.N. *Education, England and Wales: Census 2021, Usual residents aged 16 years and over who have academic, vocational, or professional qualifications, as well as the number of schoolchildren and full-time students, Census 2021 data.* 2023 10 January 2023 [cited 2024 3rd April]; Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/bulletins/educationenglandandwales/census2021#:~:text=Across%20England%20and%20Wales%2C%2033.8,18.2%25%2C%208.8%20million.
302. Statistics, A.B.o. *Education and training: Census* Information on qualifications, educational attendance and type of educational institution. 2021 28/06/2022 [cited 2024 3rd April]; Available from: https://www.abs.gov.au/statistics/people/education/education-and-training-census/2021.
303. Fok, D., R. Paap, and B. Van Dijk, *A rank-ordered logit model with unobserved heterogeneity in ranking capabilities*. Journal of applied econometrics, 2012. **27**(5): p. 831-846.
304. Craig, B., J. Busschbach, and J. Salomon, *Ranking, Time Trade-Off and Visual Analogue Scale Values for EQ-5D Health States.* Under Review, 2008.
305. Brownstone, D., *Applied Choice Analysis: A Primer*, in *Journal of the American Statistical Association*. 2007, Taylor & Francis. p. 390-390.
306. Allison, P., *What's the best R-squared for logistic regression.* Statistical Horizons, 2013. **13**.

307. Pustejovsky, J.E. and E. Tipton, *Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models.* Journal of Business & Economic Statistics, 2018. **36**(4): p. 672-683.
308. Gregory, A.W. and M.R. Veall, *Formulating Wald tests of nonlinear restrictions.* Econometrica: Journal of the Econometric Society, 1985: p. 1465-1468.
309. Rand-Hendriksen, K., et al., *Less is more: cross-validation testing of simplified nonlinear regression model specifications for EQ-5D-5L health state values.* Value in Health, 2017. **20**(7): p. 945-952.
310. Luo, N., et al., *Estimating an EQ-5D-5L value set for China.* Value in Health, 2017. **20**(4): p. 662-669.
311. Yang, Z., et al., *Cross-attribute level effects models for modeling modified 5-level version of EQ-5D health state values: is less still more?* Value in Health, 2023. **26**(6): p. 865-872.
312. Karim, S., et al., *Current Practices for Accounting for Preference Heterogeneity in Health-Related Discrete Choice Experiments: A Systematic Review.* Pharmacoeconomics, 2022. **40**(10): p. 943-956.
313. Vass, C., et al., *Accounting for preference heterogeneity in discrete-choice experiments: an ISPOR special interest group report.* Value in Health, 2022. **25**(5): p. 685-694.
314. Rungie, C.M., L.V. Coote, and J.J. Louviere, *Latent variables in discrete choice experiments.* Journal of Choice Modelling, 2012. **5**(3): p. 145-156.
315. Janssen, E.M., et al., *Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability?* Expert review of pharmacoeconomics & outcomes research, 2017. **17**(6): p. 531-542.
316. Johnson, F.R., J.-C. Yang, and S.D. Reed, *The internal validity of discrete choice experiment data: a testing tool for quantitative assessments.* Value in health, 2019. **22**(2): p. 157-160.
317. Fenwick, E.K., et al., *Development and Validation of a Preference-Based Glaucoma Utility Instrument Using Discrete Choice Experiment.* JAMA Ophthalmology, 2021. **139**(8): p. 866-874.
318. Rowen, D., et al., *Deriving a Preference-Based Measure for People With Duchenne Muscular Dystrophy From the DMD-QoL.* Value Health, 2021. **24**(10): p. 1499-1510.
319. Balestroni, G. and G. Bertolotti, *EuroQol-5D (EQ-5D): an instrument for measuring quality of life.* Monaldi Archives for Chest Disease, 2012. **78**(3).
320. Liljequist, D., B. Elfving, and K. Skavberg Roaldsen, *Intraclass correlation - A discussion and demonstration of basic features.* PLoS One, 2019. **14**(7): p. e0219854.
321. Jones, C. and J.C. Wright, *'It's time it ended and yet I hesitate, I hesitate to end it': the emotional world of an old people's home.* Journal of social work practice, 2008. **22**(3): p. 329-343.
322. Pickard, A.S., et al., *United States valuation of EQ-5D-5L health states using an international protocol.* Value in health, 2019. **22**(8): p. 931-941.
323. Luo, N., et al., *Interpretation and use of the 5-level EQ-5D response labels varied with survey language among Asians in Singapore.* Journal of Clinical Epidemiology, 2015. **68**(10): p. 1195-1204.
324. Wang, P., et al., *Valuation of EQ-5D-5L health states: a comparison of seven Asian populations.* Expert review of pharmacoeconomics & outcomes research, 2019. **19**(4): p. 445-451.

325. Cheuk Wai Ng, C., A. Wai Ling Cheung, and E. Lai Yi Wong, *Exploring potential EQ-5D bolt-on dimensions with a qualitative approach: an interview study in Hong Kong SAR, China*. Health and Quality of Life Outcomes, 2024. **22**(1): p. 42.
326. Soloman, S.R. and S.S. Sawilowsky, *Impact of rank-based normalizing transformations on the accuracy of test scores*. Journal of Modern Applied Statistical Methods, 2009. **8**(2): p. 9.
327. Himmler, S., J. van Exel, and W. Brouwer, *Estimating the monetary value of health and capability well-being applying the well-being valuation approach.* The European Journal of Health Economics, 2020. **21**: p. 1235-1244.
328. Baker, R., et al., *Public values and plurality in health priority setting: what to do when people disagree and why we should care about reasons as well as choices.* Social Science & Medicine, 2021. **277**: p. 113892.
329. Hausman, D.M., *Valuing health: Well-being, freedom, and suffering*. 2015: Oxford University Press.
330. Rogers, H.J., et al., *Discrete choice experiments or best-worst scaling? A qualitative study to determine the suitability of preference elicitation tasks in research with children and young people.* J Patient Rep Outcomes, 2021. **5**(1): p. 26.
331. Shiroiwa, T., et al., *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): p. 648-654.
332. Shiroiwa, T., et al., *Comparison of four value sets derived using different TTO and DCE approaches: application to the new region-specific PBM, AP-7D.* Health and Quality of Life Outcomes, 2024. **22**(1): p. 16.
333. Schilling, O.K., D.J. Deeg, and M. Huisman, *Affective well-being in the last years of life: The role of health decline.* Psychology and Aging, 2018. **33**(5): p. 739.
334. Jonker, M.F. and M.C. Bliemer, *On the optimization of Bayesian D-efficient discrete choice experiment designs for the estimation of QALY tariffs that are corrected for nonlinear time preferences.* Value in Health, 2019. **22**(10): p. 1162-1169.
335. Zhou, M., W.M. Thayer, and J.F. Bridges, *Using latent class analysis to model preference heterogeneity in health: a systematic review.* Pharmacoeconomics, 2018. **36**: p. 175-187.
336. Lancsar, E., et al., *Best worst discrete choice experiments in health: methods and an application.* Social science & medicine, 2013. **76**: p. 74-82.
337. Krucien, N., J. Sicsic, and M. Ryan, *For better or worse? Investigating the validity of best–worst discrete choice experiments in health.* Health Economics, 2019. **28**(4): p. 572-586.
338. Potoglou, D., et al., *Best–worst scaling vs. discrete choice experiments: an empirical comparison using social care data.* Social science & medicine, 2011. **72**(10): p. 1717-1727.
339. Burda, M., M. Harding, and J. Hausman, *A Bayesian mixed logit–probit model for multinomial choice.* Journal of econometrics, 2008. **147**(2): p. 232-246.
340. Weston, R., *Factors contributing to personal wellbeing.* Family Matters, 1999(52).
341. Al Shabasy, S., et al., *The EQ-5D-5L Valuation Study in Egypt.* Pharmacoeconomics, 2022. **40**(4): p. 433-447.
342. Andrade, L.F., et al., *A French Value Set for the EQ-5D-5L.* Pharmacoeconomics, 2020. **38**(4): p. 413-425.
343. Bouckaert, N., et al., *An EQ-5D-5L Value Set for Belgium.* Pharmacoecon Open, 2022: p. 1-14.

344. Chemli, J., et al., *Valuing health-related quality of life using a hybrid approach: Tunisian value set for the EQ-5D-3L.* Qual Life Res, 2021.
345. Chen, G., et al., *Quality of care experience in aged care: An Australia-Wide discrete choice experiment to elicit preference weights.* Soc Sci Med, 2021. **289**: p. 114440.
346. Ferreira, P.L., et al., *A hybrid modelling approach for eliciting health state preferences: the Portuguese EQ-5D-5L value set.* Qual Life Res, 2019. **28**(12): p. 3163-3175.
347. Finch, A.P., et al., *An EQ-5D-5L value set for Italy using videoconferencing interviews and feasibility of a new mode of administration.* Soc Sci Med, 2022. **292**: p. 114519.
348. Finch, A.P., et al., *Estimation of an EORTC QLU-C10 Value Set for Spain Using a Discrete Choice Experiment.* Pharmacoeconomics, 2021. **39**(9): p. 1085-1098.
349. Gamper, E.M., et al., *EORTC QLU-C10D value sets for Austria, Italy, and Poland.* Qual Life Res, 2020. **29**(9): p. 2485-2495.
350. Gutierrez-Delgado, C., et al., *EQ-5D-5L Health-State Values for the Mexican Population.* Appl Health Econ Health Policy, 2021. **19**(6): p. 905-914.
351. Hansen, T.M., K. Stavem, and K. Rand, *Sample Size and Model Prediction Accuracy in EQ-5D-5L Valuations Studies: Expected Out-of-Sample Accuracy Based on Resampling with Different Sample Sizes and Alternative Model Specifications.* MDM Policy Pract, 2022. **7**(1): p. 23814683221083839.
352. Jansen, F., et al., *Dutch utility weights for the EORTC cancer-specific utility instrument: the Dutch EORTC QLU-C10D.* Qual Life Res, 2021.
353. Jensen, C.E., et al., *The Danish EQ-5D-5L Value Set: A Hybrid Model Using cTTO and DCE Data.* Appl Health Econ Health Policy, 2021.
354. Jiang, E.X., et al., *Calculating Ex-ante Utilities From the Modified Japanese Orthopedic Association Score: A Prerequisite for Quantifying the Value of Care for Cervical Myelopathy.* Spine (Phila Pa 1976), 2022. **47**(7): p. 523-530.
355. Kemmler, G., et al., *German value sets for the EORTC QLU-C10D, a cancer-specific utility instrument based on the EORTC QLQ-C30.* Qual Life Res, 2019. **28**(12): p. 3197-3211.
356. King, M.T., et al., *The Functional Assessment of Cancer Therapy Eight Dimension (FACT-8D), a Multi-Attribute Utility Instrument Derived From the Cancer-Specific FACT-General (FACT-G) Quality of Life Questionnaire: Development and Australian Value Set.* Value Health, 2021. **24**(6): p. 862-873.
357. Kreimeier, S., et al., *EQ-5D-Y Value Set for Germany.* Pharmacoeconomics, 2022: p. 1-13.
358. Ludwig, K., V.D.S.J. Graf, and W. Greiner, *German Value Set for the EQ-5D-5L.* Pharmacoeconomics, 2018. **36**(6): p. 663-674.
359. McTaggart-Cowan, H., et al., *The EORTC QLU-C10D: The Canadian Valuation Study and Algorithm to Derive Cancer-Specific Utilities From the EORTC QLQ-C30.* MDM Policy Pract, 2019. **4**(1): p. 2381468319842532.
360. Miguel, R.T.D., et al., *Estimating the EQ-5D-5L value set for the Philippines.* Qual Life Res, 2022. **31**(9): p. 2763-2774.
361. Nerich, V., et al., *French Value-Set of the QLU-C10D, a Cancer-Specific Utility Measure Derived from the QLQ-C30.* Appl Health Econ Health Policy, 2021. **19**(2): p. 191-202.
362. Omelyanovskiy, V., et al., *Valuation of the EQ-5D-3L in Russia.* Qual Life Res, 2021.
363. Pattanaphesaj, J., et al., *The EQ-5D-5L Valuation study in Thailand.* Expert Rev Pharmacoecon Outcomes Res, 2018. **18**(5): p. 551-558.

364. Pahuta, M.A., et al., *Calculating Utilities From the Spine Oncology Study Group Outcomes Questionnaire: A Necessity for Economic and Decision Analysis.* Spine (Phila Pa 1976), 2021. **46**(17): p. 1165-1171.
365. Pickard, A.S., et al., *United States Valuation of EQ-5D-5L Health States Using an International Protocol.* Value Health, 2019. **22**(8): p. 931-941.
366. Prevolnik, R.V. and M. Ogorevc, *EQ-5D-Y Value Set for Slovenia.* Pharmacoeconomics, 2021.
367. Ramos-Goñi, J.M., et al., *Accounting for Unobservable Preference Heterogeneity and Evaluating Alternative Anchoring Approaches to Estimate Country-Specific EQ-5D-Y Value Sets: A Case Study Using Spanish Preference Data.* Value Health, 2022. **25**(5): p. 835-843.
368. Ramos-Goñi, J.M., et al., *Does Changing the Age of a Child to be Considered in 3-Level Version of EQ-5D-Y Discrete Choice Experiment-Based Valuation Studies Affect Health Preferences?* Value Health, 2022.
369. Rencz, F., et al., *Value Set for the EQ-5D-Y-3L in Hungary.* Pharmacoeconomics, 2022: p. 1-11.
370. Revicki, D.A., et al., *United States Utility Algorithm for the EORTC QLU-C10D, a Multiattribute Utility Instrument Based on a Cancer-Specific Quality-of-Life Instrument.* Med Decis Making, 2021. **41**(4): p. 485-501.
371. Rogers, H.J., et al., *Adolescent valuation of CARIES-QC-U: a child-centred preference-based measure of dental caries.* Health Qual Life Outcomes, 2022. **20**(1): p. 18.
372. Roudijk, B., et al., *A Value Set for the EQ-5D-Y-3L in the Netherlands.* Pharmacoeconomics, 2022: p. 1-11.
373. Tsuchiya, A., et al., *Manipulating the 5 Dimensions of the EuroQol Instrument: The Effects on Self-Reporting Actual Health and Valuing Hypothetical Health States.* Med Decis Making, 2019. **39**(4): p. 379-392.
374. Webb, E., et al., *Transforming discrete choice experiment latent scale values for EQ-5D-3L using the visual analogue scale.* Eur J Health Econ, 2020. **21**(5): p. 787-800.
375. Welie, A.G., et al., *Valuing Health State: An EQ-5D-5L Value Set for Ethiopians.* Value Health Reg Issues, 2020. **22**: p. 7-14.
376. Xie, S., et al., *Do Discrete Choice Experiments Approaches Perform Better Than Time Trade-Off in Eliciting Health State Utilities? Evidence From SF-6Dv2 in China.* Value Health, 2020. **23**(10): p. 1391-1399.
377. Wentao Zhu et al., *CQ-11D measure explanation.* China Pharma Eco, 2022. **17**(5): p. 16-20,45.