Development of a value-based scoring system for the WAItE using the OPUF in a sample of adults

Will King Tomos Robinson Angela Bate Laura Ternent

Background: Online personal utility functions (OPUF) present a new method for eliciting preferences. In this study we used the OPUF to elicit a health state utility value set for the Weight-specific Adolescent Instrument for Economic-evaluation (WAItE) with a representative sample of the UK adult population.

Methods: WAItE OPUF survey design was informed by prior qualitative work. The survey consisted of the WAItE descriptive system, domain weighting, level rating (per domain) and a VAS anchoring task. Personal utility functions were estimated on the individual level for all participants. Personal utility functions were aggregated and combined with the anchoring factor to give the social utility function and utility value set. Preference heterogeneity was assessed using Euclidean distance and PERMANOVA to explore preference variation within the sample. An experimental sensitivity analysis dichotomised preference heterogeneity into anchoring variation and domain weighting/level rating variation.

Results: A total of 300 participants completed the WAItE OPUF survey. The sample was broadly representative of the UK adult population. Participants, on average, took less than 10 minutes to complete the survey. The most important domains were tiredness and unhappiness, while least important domains were sports and embarrassment. Social utility values and the anchoring utility value estimated were comparable to previous studies. Preferences generally were heterogeneous, especially among different ages. Younger participants assigned lower utility values to WAItE health states and provided significantly lower scores on the VAS anchoring task compared to older participants.

Conclusion: This study successfully elicited health state utility values for the WAItE using the OPUF. Preference heterogeneity analysis identified differences in preferences for different age groups and a further valuation study is ongoing to explore whether this heterogeneity exists between adults and adolescents.

Introduction

This chapter presents the introduction, methods, results and discussion from an empirical study developing a utility value set for the WAItE using online personal utility functions (OPUF) with a representative sample of UK adults.

Compositional preference elicitation methods

Preference elicitation methods generally speaking, fall into two categories: compositional and decompositional (Keeney and Raiffa 1979; Marsh et al. 2016; Belton and Stewart 2002). That is, methods like DCE, BWS and TTO elicit preference orderings from individuals for an entire health state (composed of a combination of domains and levels) and then responses are decomposed to identify marginal contributions of each domain and level in each health state. Models like multinomial logit, mixed logit and latent class are frequently used to decompose responses to decompositional preference elicitation tasks (Hauber et al. 2016). Coefficients estimated in these models form the basis of dis/utility values for each domain and level in a descriptive system.

Conversely, compositional methods seek to identify preferences for each domain weighting and level rating individually for the number of domains and levels in a given descriptive system. Therefore, statistical models to elicit coefficients for each individual domain and level are not required and responses to each domain weighting and level rating are combined (in addition to an anchoring factor) to yield dis/utility values for each domain and level in the descriptive system. Compositional approaches can take many forms from simple VAS scores to using semantic categories and ranking methods (Bana E Costa and Vansnick 1999; Danner et al. 2011; Oliveira et al. 2018). These approaches have been used successfully in multi-criteria decision analysis (MCDA), but have been used less extensively in the preference elicitation space. Since the development of the OPUF, compositional approaches to elicit preferences have become more commonplace and a number of countries are using the OPUF to elicit value sets specific to their population (Brodszky et al. 2023).

From PUF to OPUF

Personal utility functions were first used in the context of preference elicitation by Devlin et al. (2019) (Devlin et al. 2019) to estimate the feasibility for using this approach to estimate a value set for the EQ5D-5L. Since the feasibility for the underlying PUF methods were established, the approach has been expanded by Schneider and colleagues and converted into an online personal utility functions (OPUF) survey built initially using RShiny (P. P. Schneider et al. 2022) and subsequently using Javascript (available here). Since the development of the OPUF, a number of descriptive systems and different research teams have begun utilising

this method to elicit value sets (Bray, Tudor Edwards, and Schneider 2024; Brodszky et al. 2023).

An overview of the OPUF structure

1. Domain weighting: This section is composed of two parts. First, domain ranking is completed where participants identify their most important domain. Second, respondents complete the domain weighting (swing weighting) where the relative importance of other domains is ascertained using their most important domain as a reference point. These questions are presented in Figure **??**.
2. Level ratings: This element of the OPUF has varied across different iterations of the survey. Schnieder et al. (2022) (P. P. Schneider et al. 2022) asked participants to rank the levels within the descriptive system generally (i.e. for any given domain), while other iterations have administered separate level rating questions for each domain in the descriptive system (Bray, Tudor Edwards, and Schneider 2024). Selection of method requires a trade-off between participant burden and sensitivity of level ratings to each domain. Figure **??** presents the level rating question for the pain/discomfort domain of the EQ5D-5L.
3. Anchoring factor: A task is required to rescale the latent coefficients estimated via combining level ratings and domain weights onto the QALY scale. Participants are presented with a binary choice between the PITS state (or another state) of a given descriptive system and "being dead". If the PITS state is chosen, participants are asked to rank the PITS state on a VAS from 1 (full health) to 0 (dead). If "being dead" is chosen, participants are asked to rank "being dead" on a VAS from 1 (full health) to 0 (PITS state). Responses to these respective questions provide the anchoring factor. Anchoring questions, such that PITS is preferred to dead, are presented in Figure **??**.

OPUF logic and mathematics

This section presents the logic and underlying mathematics required to convert the raw OPUF responses from one person into an anchored value set for the WAItE descriptive system. This example assumes that level ratings are obtained for each domain separately, therefore mathematics presented here differs to those presented elsewhere (P. P. Schneider et al. 2022). Example response data are used for demonstration in this section and are presented in Table 1.

Example responses

Table 1: Example individual responses to the OPUF

Response	Tired	Walking	Sports	Concentration	Embarrassment	Unhappiness	Tr
Level rating: Never	0	0	0	0	0	0	
Level rating: Almost Never	14	26	21	15	16	12	
Level rating: Sometimes	57	55	63	54	38	26	
Level rating: Often	83	82	85	86	64	38	
Level rating: Always	100	100	100	100	100	100	
Domain Weighting	28	33	36	45	100	34	

Note: WAItE PITS better than dead = Yes Anchoring Task Response = 20 PITS Utility Value = 0.2

Level ratings (presented in Table 1) are converted to coefficients bounded between 0-1 (shown in Equation 1). Level rating coefficients are presented in Equation 2. Attribute weights (presented in Table 1) are then normalised to sum to the value of 1 by dividing each weight by the sum of all weights (shown in Equation 3). Normalised attribute weights are presented in Equation 4.

$$L_{ij} \cdot 0.01 \tag{1}$$

$$L_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.14 & 0.26 & 0.21 & 0.15 & 0.16 & 0.12 & 0.19 \\ 0.57 & 0.55 & 0.63 & 0.54 & 0.38 & 0.26 & 0.66 \\ 0.83 & 0.82 & 0.85 & 0.86 & 0.64 & 0.38 & 0.91 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{2}$$

$$\frac{w_j}{\sum w_j} \tag{3}$$

$$w_j = \begin{bmatrix} 0.08 & 0.10 & 0.11 & 0.14 & 0.30 & 0.10 & 0.17 \end{bmatrix} \tag{4}$$

Combining the attribute weights (Equation 4) with the level coefficients (Equation 2) via element-wise multiplication (shown in Equation 5 gives the coefficient matrix presented in Equation 6}. Once the coefficient matrix has been estimated, preference values can be estimated on the 0-1 QALY scale where the worst health state (PITS state denoted 5555555) is zero and the best health state (denoted 1111111) is one. These latent coefficients must now be rescaled to incorporate the results from the PITS anchoring task so that the minimum utility value possible is equal to the PITS value.

$$L_{ij} \cdot w_j = \tilde{M}_{ij} \tag{5}$$

$$\tilde{M}_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0.03 & 0.02 & 0.02 & 0.05 & 0.01 & 0.03 \\ 0.05 & 0.05 & 0.07 & 0.07 & 0.11 & 0.03 & 0.11 \\ 0.07 & 0.08 & 0.09 & 0.12 & 0.19 & 0.04 & 0.15 \\ 0.08 & 0.10 & 0.11 & 0.14 & 0.30 & 0.10 & 0.17 \end{bmatrix} \tag{6}$$

To rescale the latent coefficient matrix to incorporate the anchoring task, the coefficient matrix is multiplied by the compliment of the PITS value (shown in Equation 7) to give the anchored coefficient matrix presented in Equation 8.

$$\tilde{M}_{ij} \cdot (1 - P) \qquad P = 0.2 \tag{7}$$

$$\tilde{V}_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.01 & 0.02 & 0.02 & 0.02 & 0.04 & 0.01 & 0.02 \\ 0.04 & 0.04 & 0.06 & 0.06 & 0.09 & 0.02 & 0.09 \\ 0.06 & 0.06 & 0.07 & 0.10 & 0.15 & 0.03 & 0.12 \\ 0.06 & 0.08 & 0.09 & 0.11 & 0.24 & 0.08 & 0.14 \end{bmatrix} \tag{8}$$

Once the attribute and level labels are reintroduced to the anchored coefficient matrix this forms the value set which presents the disutility corresponding to each attribute level combination presented in the WAItE. Table 2 presents the WAItE example PUF value set. Equation 9 present examples of how to estimate a utility value given a specific WAItE health state.

Table 2: WAItE example PUF value set

Table 2: WAItE example PUF value set

X.	Tired	Walking	Sports	Concentration	Embarrassment	Unhappiness	Treated.Differently
Never	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Almost Never	0.01	0.02	0.02	0.02	0.04	0.01	0.02
Sometimes	0.04	0.04	0.06	0.06	0.09	0.02	0.09
Often	0.06	0.06	0.07	0.10	0.15	0.03	0.12
Always	0.06	0.08	0.09	0.11	0.24	0.08	0.14

$$\text{Health State } [5555555] \Rightarrow 1 - (0.06 + 0.08 + 0.09 + 0.11 + 0.24 + 0.08 + 0.14) = 0.20 \text{Health State } [5223445] \quad \Rightarrow \tag{9}$$

Table 3: WAItE example PUF value set

	Tired	Walking	Sports	Concentration	Embarrassment	Unhappiness	Treated Differently
Never	0	0	0	0	0	0	0
Almost Never	0.01	0.02	0.02	0.02	0.04	0.01	0.02
Sometimes	0.04	0.04	0.06	0.06	0.09	0.02	0.09
Often	0.06	0.06	0.07	0.10	0.15	0.03	0.12
Always	0.06	0.08	0.09	0.11	0.24	0.08	0.14

*Coefficients anchored by a PITS utility value of 0.2

Aggregation to social utility function

The OPUF is designed to be able to estimate personal utility functions and so estimation occurs on an individual basis. Aggregating personal utility functions to a social utility function (SUF) takes place by taking a mean of all the individual personal utility functions from your sample. This operation is presented in Equation 10.

$$\bar{V}_{ij} = \frac{\sum_{\tilde{V}_{ij}}}{N} \tag{10}$$

Methods

Recruitment

This study recruited (n=300) adults to respond to a quality-of-life survey hosted online. Study participants were recruited based on specific quotas to form a representative sample based on UK census data. The survey was hosted on the Prolific platform which invited paid respondents to complete the Weight-specific Adolescent Instrument for Economic evaluation (WAItE) version of the Online Personal Utility Functions (OPUF) survey. A demonstration of the OPUF survey and questions is available here. Informed consent was obtained at the outset of the survey and participants reserved the right to withdraw at any point without giving a reason. Participants who withdrew were not paid and their data deleted. Participation in this survey was estimated to take approximately fifteen minutes to complete and participants received £2.50 as a payment upon completion. This is in line with reimbursements rates from other OPUF studies (P. P. Schneider et al. 2022; Bray, Tudor Edwards, and Schneider 2024) and is in line with recommended reimbursement rates from Prolific (www.prolific.com). The survey was designed to be an unassisted survey administered online (no face-to-face contact) and no identifiable data was collected. Statistical analysis was conducted on the survey data. Newcastle University Medical School Ethics Committee approved this study (reference 49737/2023). The survey structure is detailed in Section .

Survey structure

1. Consent and Prolific ID: Participants were asked to consent to participate and enter their unique Prolific ID. This enables demographic information held by Prolific on their participants to be linked to each respondent.
2. WAItE descriptive system: Participants were asked to complete the WAItE descriptive system (presented in Figure **??**)to describe their current health state.
3. Dimension selection}: Participants were presented with the worst level for each WAItE dimension and asked to choose which health problem would have the most negative impact on their quality of life. The dimension chosen is then used in the subsequent dimension swing weighting task.

4. Dimension swing weighting: Participants were presented with each dimension in the WAItE and asked to consider an improvement from the worst level of that dimension to the best level of that dimension. Participants were asked to rank this improvement on a visual slider from 0-100 where the most important dimension (chosen in the previous task) is fixed at 100. Participants were reminded to use their most important dimension as a reference point.

5. Level rating: Participants were presented with a specific dimension of the WAItE and shown each level within that dimension. Levels best and worst (never and always) were fixed at 0 and 100 respectively. Participants were asked to rank the intermediate levels within each dimension using the fixed levels as a reference point.
6. Anchoring: Participants were presented with a binary choice asking whether they prefer the worst state of the WAItE (PITS state) or death. If participants choose the worst state of the WAItE, a second question is asked which asks them to rank the WAITE PITS state on a visual analogue scale where zero is labelled as being dead and one hundred is labelled as no health problems. If participants choose death in the binary choice, they were asked to rank being dead on a visual analogue scale where zero is labelled as the WAItE PITS state and one hundred is labelled as no health problems.
7. Survey feedback and demographic questions: Participants were asked about how difficult they found the task to complete and demographic information on age, gender, ethnicity, education, employment and weight status.

Live survey

Missing data

Through the survey design process the potential for large amounts of missing data has been mitigated by ensuring responses were compulsory to certain questions. However for ethical reasons, we allowed participants to not answer the questions relating to death. For participants who do not provide responses to the anchoring questions, their responses were imputed using

multiple imputation by chained equations (MICE) (White, Royston, and Wood 2011) which were informed by demographic information and dimension weighting responses.

Preference heterogeneity

As personal utility function are estimated on an individual basis, exploring preference heterogeneity between individuals in the sample is straightforward. Investigating the heterogeneity of preferences between individuals, requires a measure of dis/similarity to quantify how far apart two PUFs are (P. Schneider et al. 2024). The measurement and estimation of preference heterogeneity in this section will follow methods detailed by Schneider et al. (2024) (P. Schneider et al. 2024). Each PUF estimated in this study was represented by a vector of 78,125 health state utility values for each respondent in the sample. In order to assess the dis/similarity between these PUFs, we used the euclidean difference measure (EUD). Analogous to a line between two points on a two dimensional plane, the EUD between two PUFs denotes the shortest path length in a 78,125 dimensional space. It is computed as the square root of the sum of the squared differences between the PUFs of individuals (i) and (j) (presented in Equation 11). Once PUFs have been estimated for all individuals in the sample, pairwise EUD was estimated for all possible pairwise combinations within the sample. Pairwise EUD was stored in an [N (×) N] distance matrix.

$$d_{EUD}(i,j) = \sqrt{\sum (u_i(s_1) - u_j(s_1))^2 + ... + (u_i(s_{78125}) - u_j(s_{78125}))^2} \\ s = \{1111111, 2111111, ..., 5555555\} \tag{11}$$

Permutational analysis of variance

Permutational analysis of variance (PERMANOVA), analogous to analysis of variance, is a geometric partitioning of variation across a multivariate data cloud, definied in the space of any given dissimilarity measure, in response to one or more groups (Anderson 2017; Anderson and Walsh 2013). This method of statistical testing has been used most commonly in ecological research to test for population dispersion among different subgroups (Souza et al. 2013). PERMANOVA decomposes the total distances between observations (SS(_T)) into within-groups (SS(_W)) and between groups sum-of-squares (SS(_B)). Equation 12 details the estimation of total and within-groups sum-of-squares. Mathematical notation presented here is reproduced from Schneider et al. (2024) (P. Schneider et al. 2024) for consistency.

$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d(i,j)^2; \quad SS_W = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d(i,j)^2 \epsilon_{ij}^{\ell} / n_{\ell} \tag{12}$$

where N is the total sample size (=300), (d(i,j)^2) is the squared distance between the PUFs of participants (i) and (j), (*{i,j}) indicator which is 1, if participants (i) and (j) belong to the same group, and 0 if they do not, and (n{ }) is the size for group ( ). Then, SS(_B) can then be calculated as SS(_B) = SS(T) – SS(_W), which allows calculating the pseudo F statistic for (p) groups:

$$F = \frac{(\frac{SS_B}{p-1})}{(\frac{SS_W}{N-p})} \tag{13}$$

Further details about the mathematical and statistical properties of PERMANOVA are available elsewhere [P. Schneider et al. (2024); Anderson2017; Anderson2013PERMANOVATesting]. In this study, we used PERMANOVA to explore the variability in WAItE health state preferences (individual value sets) between various subgroups. A multivariate PERMANOVA model was estimated with subgroups of: age, gender, self-reported weight status, education, employment status and ethncity.

Sensitivity analysis

In an experimental sensitivity analysis, preference heterogeneity was assessed using EUD estimated based on individual's personal utility functions anchored using the social PITS utility value (henceforth referred to as EUD2). This differed to prior preference heterogeneity estimation as individual variation in PITS utility values were not included in the EUD2 estimation. EUD2 was entirely composed by differences in level ratings and domain weights. Further details on the derivation of EUD2 are presented in Appendix **??**.

Results

Study participants

A sample of 334 individuals were approached to participate in the study via the survey company Prolific. Individuals that successfully inputted their unique Prolific ID and obtained a correct completion code from the end of the study were included in the analysis sample and received a small payment (£2.50) for their participation. Seven participants were excluded from the study as they had an incorrect completion code and did not enter the correct unique Prolific ID. Therefore no data was available on those seven participants and so they were excluded from the analysis. An additional participant was excluded from the analysis due to completing the survey in eighteen seconds (well under the prespecified minimum time limit of 2 minutes). Two respondents timed-out while completing the survey and were therefore not included. Twenty-four individuals chose not complete the study (referred to by Prolific as 'returned' participants). This left an analysis sample of N=300 participants who successfully completed the survey. A

representative sample based on UK census data was obtained from Prolific. A summary of demographic information collected in the OPUF are presented in Table **??**.

Survey duration

The mean (SD) and median (IQR) survey completion time in minutes was 9.66 (5.85) and 8.15 (5.88; 11.89). Table 4 summarises how much time was spent completing each individual section of the survey.

Table 4: Survey completion times (secs)

Section	Mean (SD)	Median (Q1; Q3)	Min	Max
WAItE	73 (90.1)	53 (38; 77)	10	1066
Dimension ranking	35.4 (50.6)	26 (16; 40)	2	741
Dimension weighting	115.7 (94.9)	91.5 (69.8; 142.2)	18	1380
Level rating	220.5 (206.4)	171 (119; 249)	34	2158
PITS vs death	25.5 (45.4)	16.5 (11; 25.2)	4	620
PITS-VAS	37.3 (39.2)	29 (21; 45)	5	605
PITS-VAS	37.3 (39.2)	29 (21; 45)	5	605
Total (secs)	579.5 (351.1)	489.2 (352.5; 713.4)	126.7	3738.2
Total (mins)	9.66 (5.85)	8.15 (5.88; 11.89)	2.11	62.3

WAItE descriptive system

Responses to the WAItE descriptive system are presented in Table **??**. Feeling tired and avoiding doing sport were the domains that were most frequently experienced by participants in our analysis sample. WAItE summary statistics were in line with results from previous studies (Robinson and Oluboyede 2019).

Level ratings

Level ratings are presented individually for each different domain in Table 5. The best and worst levels (*Always* and *Never*) were fixed at 0 and 100 respectively. The second best level (*Almost never*) had the lowest VAS score in the Sports and Embarrassment domain, while the second worst level (*Often*) had the highest VAS score in the Concentration domain. In this question, higher VAS scores indicate worse states of health.

Table 5: Summary of OPUF level ratings by domain

Section	Mean (SD)	Median (Q1; Q3)	Min	Max
Almost never	20.323 (23.208)	10 (5; 25)	0	100
Sometimes	36.31 (19.185)	33.5 (20; 50)	0	100
Often	62.217 (23.934)	70 (50; 80)	0	100
Almost never	19.39 (21.839)	10 (6; 21)	0	100
Sometimes	37.677 (19.373)	40 (24; 50)	0	100
Often	62.967 (26.167)	71 (50; 80)	0	100
Almost never	16.63 (20.978)	10 (5; 20)	0	100
Sometimes	29.487 (22.015)	25 (10; 45)	0	100
Often	49.843 (29.624)	50.5 (24.5; 75)	0	100
Almost never	21.393 (22.1)	14 (7; 25)	0	100
Sometimes	41.56 (20.101)	40 (25.8; 53.2)	0	100
Often	64.503 (26.195)	73 (50; 80.2)	0	100
Almost never	16.59 (22.292)	10 (4; 20)	0	100
Sometimes	29.417 (21.615)	25 (10; 50)	0	100
Often	47.91 (30.445)	50 (20; 75)	0	100
Almost never	21.13 (22.235)	13 (6; 25)	0	100
Sometimes	41.363 (22.128)	41.5 (25; 56)	0	100
Often	63.557 (28.187)	75 (50; 85)	0	100
Almost never	20.93 (24.366)	11 (5; 25)	0	100
Sometimes	35.52 (22.774)	34.5 (19.8; 50)	0	100
Often	55.857 (30.552)	60.5 (31; 80)	0	100

Domain weights

Summary statistics of domain weightings are presented in Table 6. On average, Tiredness (76.5) and Unhappiness (70) were considered to be more important to participants than Embarrassment (40.1) and Sports (42.3). There was less variability in domain weighting responses to Tiredness than responses to Treated differently or Embarrassment. %TODO: Add relative attribute importance section with normalised RAI score

Table 6: Summary of OPUF domain weights and anchoring responses

Section	Mean (SD)	Median (Q1; Q3)	Min	Max
Tired	76.513 (28.358)	90 (60; 100)	1	100
Walking	65.53 (32.49)	75 (40; 100)	0	100
Sports	42.32 (32.81)	35 (11; 70)	0	100
Concentration	67.897 (30.949)	80 (44; 99.2)	0	100
Embarrassment	40.143 (34.344)	30 (9; 70)	0	100

Table 6: Summary of OPUF domain weights and anchoring responses

Section	Mean (SD)	Median (Q1; Q3)	Min	Max
Unhappiness	69.997 (31.946)	80 (50; 100)	0	100
Treated differently	52.093 (35.564)	50 (15.8; 86)	0	100
PITS preferred to death	0.879 (0.327)	1 (1; 1)	0	1
PITS-VAS	56.057 (31.287)	54 (30; 85)	0	100
Dead-VAS	42.528 (31.583)	38.5 (13.2; 63.5)	1	100
PITS VAS uncensored	-0.025 (5.95)	0.5 (0.2; 0.8)	-99	1
PITS VAS censored	0.431 (0.485)	0.5 (0.2; 0.8)	-1	1
PITS Utility Value	0.282 (1.456)	0.5 (0.2; 0.8)	-14.3	1

Anchoring

The majority of respondents in the sample preferred the WAItE PITS state to being dead (87%). Therefore, 13% of participants answered the dead-VAS and 87% answered the PITS-VAS. A proportion of participants did not answer the anchoring task (1.67%). After winsorizing extreme values (top and bottom 0.1%) (*Applying Contemporary Statistical Techniques* 2003) and conducting multiple imputation by chained equations on the missing values, the mean (SD) and median (IQR) PITS utility value was 0.282 (1.456) and 0.5 (0.6). The distribution of WAItE PITS utility values (after winsorizing and imputation) is presented in Figure 1.

Social utility function estimation

Personal utility functions were estimated individually for each participant in our analysis sample via methods outlined in Section . After this, individual PUFs were aggregated into a group utility function and anchored using the group PITS utility value (0.282) to give the social utility function. Descriptive statistics from the social utility function are presented in Table 7 whereby the mean values can be used to estimate utility values for WAItE health states.

Table 7: Social utility function based on 300 PUFs

Dimension Level	Mean (95% CI)	Median (Q1; Q3)	Min	Max
tired, almost never	0.029 (0.025; 0.033)	0.029 (0.027; 0.03)	0.021	0.039
tired, sometimes	0.052 (0.048; 0.057)	0.052 (0.05; 0.054)	0.042	0.062
tired, often	0.088 (0.082; 0.094)	0.088 (0.085; 0.09)	0.077	0.103
tired, always	0.14 (0.133; 0.148)	0.14 (0.137; 0.143)	0.125	0.157
walk, almost never	0.021 (0.018; 0.024)	0.021 (0.02; 0.022)	0.016	0.027
walk, sometimes	0.045 (0.041; 0.049)	0.045 (0.043; 0.046)	0.037	0.052

Table 7: Social utility function based on 300 PUFs

Dimension Level	Mean (95% CI)	Median (Q1; Q3)	Min	Max
walk, often	0.075 (0.069; 0.082)	0.075 (0.073; 0.077)	0.064	0.087
walk, always	0.116 (0.108; 0.124)	0.115 (0.113; 0.118)	0.101	0.131
sports, almost never	0.012 (0.01; 0.015)	0.012 (0.012; 0.013)	0.009	0.017
sports, sometimes	0.023 (0.02; 0.025)	0.023 (0.022; 0.024)	0.018	0.029
sports, often	0.038 (0.034; 0.044)	0.038 (0.037; 0.04)	0.029	0.052
sports, always	0.069 (0.063; 0.076)	0.069 (0.067; 0.071)	0.058	0.084
concentrate, almost never	0.026 (0.023; 0.03)	0.026 (0.025; 0.028)	0.019	0.034
concentrate, sometimes	0.051 (0.047; 0.055)	0.051 (0.049; 0.052)	0.044	0.06
concentrate, often	0.08 (0.074; 0.086)	0.08 (0.078; 0.082)	0.069	0.093
concentrate, always	0.121 (0.114; 0.128)	0.121 (0.118; 0.123)	0.107	0.138
embarrassment, almost never	0.012 (0.01; 0.014)	0.012 (0.011; 0.013)	0.008	0.017
embarrassment, sometimes	0.022 (0.019; 0.025)	0.022 (0.021; 0.023)	0.016	0.027
embarrassment, often	0.034 (0.03; 0.038)	0.034 (0.032; 0.035)	0.025	0.043
embarrassment, always	0.061 (0.056; 0.067)	0.061 (0.059; 0.063)	0.051	0.072
unhappiness, almost never	0.025 (0.022; 0.029)	0.025 (0.024; 0.026)	0.019	0.033
unhappiness, sometimes	0.054 (0.049; 0.059)	0.054 (0.052; 0.056)	0.045	0.064
unhappiness, often	0.083 (0.076; 0.09)	0.083 (0.081; 0.086)	0.07	0.101
unhappiness, always	0.124 (0.117; 0.133)	0.124 (0.122; 0.127)	0.11	0.142
treated differently, almost never	0.019 (0.016; 0.022)	0.019 (0.017; 0.02)	0.013	0.025
treated differently, sometimes	0.035 (0.03; 0.039)	0.035 (0.033; 0.036)	0.026	0.043
treated differently, often	0.052 (0.047; 0.058)	0.052 (0.05; 0.054)	0.041	0.063
treated differently, always	0.087 (0.079; 0.095)	0.087 (0.084; 0.089)	0.071	0.101

Figure 2 presents the mean social utility function (thick line) alongside individual personal utility functions (thin lines) for a selection of 100 WAItE health states ordered from high to low utility according to the social preference. Deviations of individual utility functions from the social preference illustrate the heterogeneity of preference within our analysis sample. Individual personal utility functions shown in Figure 2 are anchored using individual PITS utility values rather than the social PITS utility value.

Preference heterogeneity

After estimating individual PUFs for all participants, pairwise EUD was estimated between all participants. This yielded a [300 (×) 300] distance matrix with 44,850 unique pairwise comparisons. The mean (SD) and median (IQR) EUD were 47.60 (19.40) and 44.81 (34.06; 57.40). The highest and lowest observed EUD were 189.16 and 0. Figure 3 illustrates the relationship between EUD and WAItE health states. EUD tends to increase as WAItE health

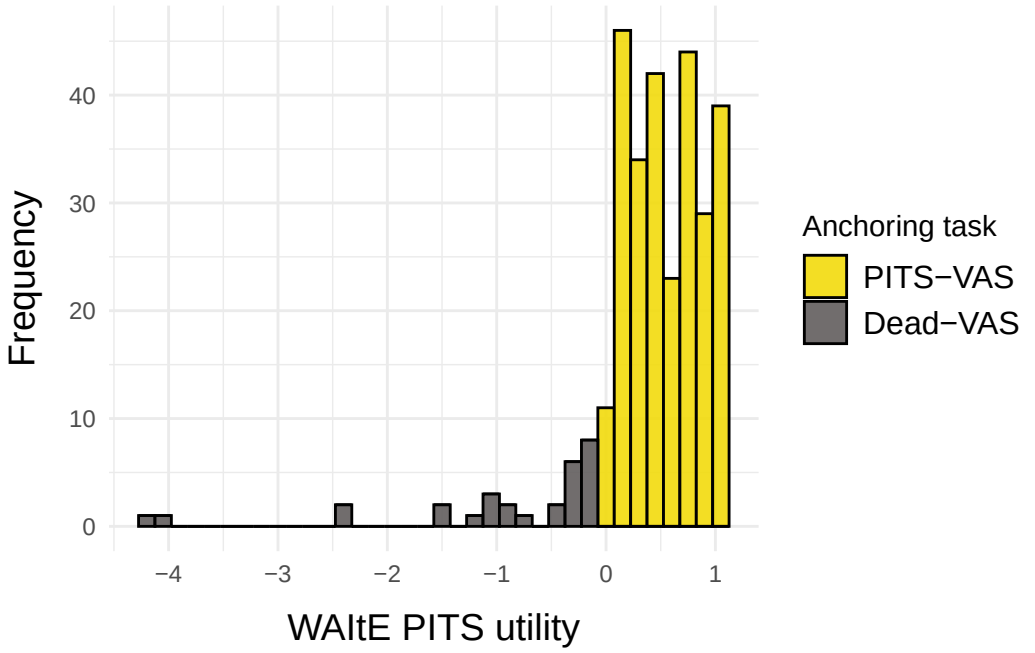

Figure 1: Distribution of PITS utility values

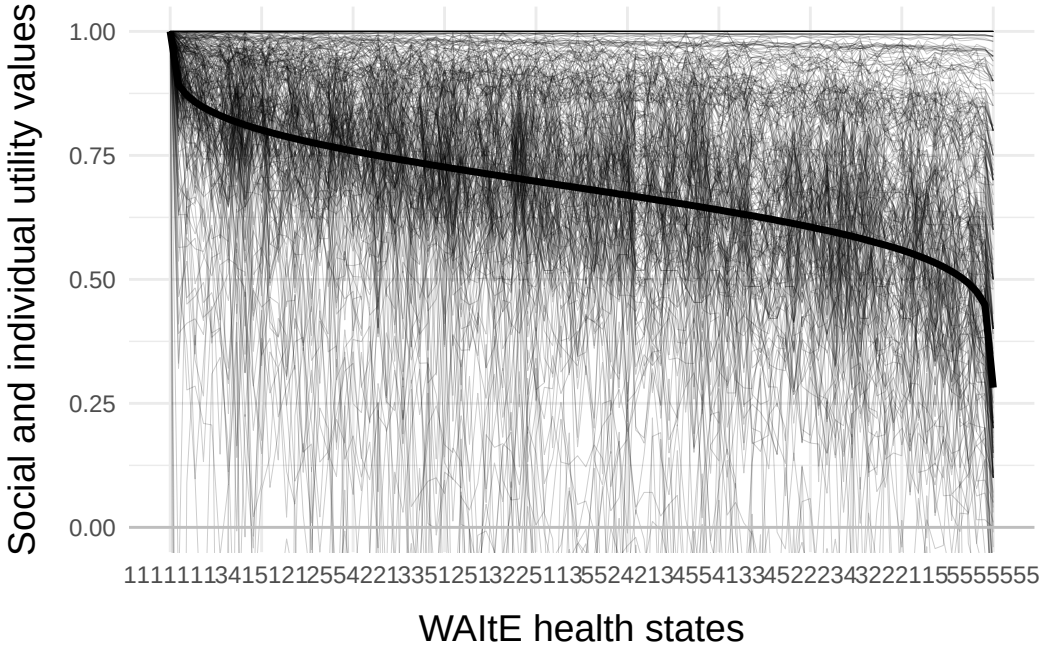

Figure 2: Social and individual utility functions

states worsen. That is, as the severity of WAItE health states increases, the more heterogeneous preferences become among our sample.

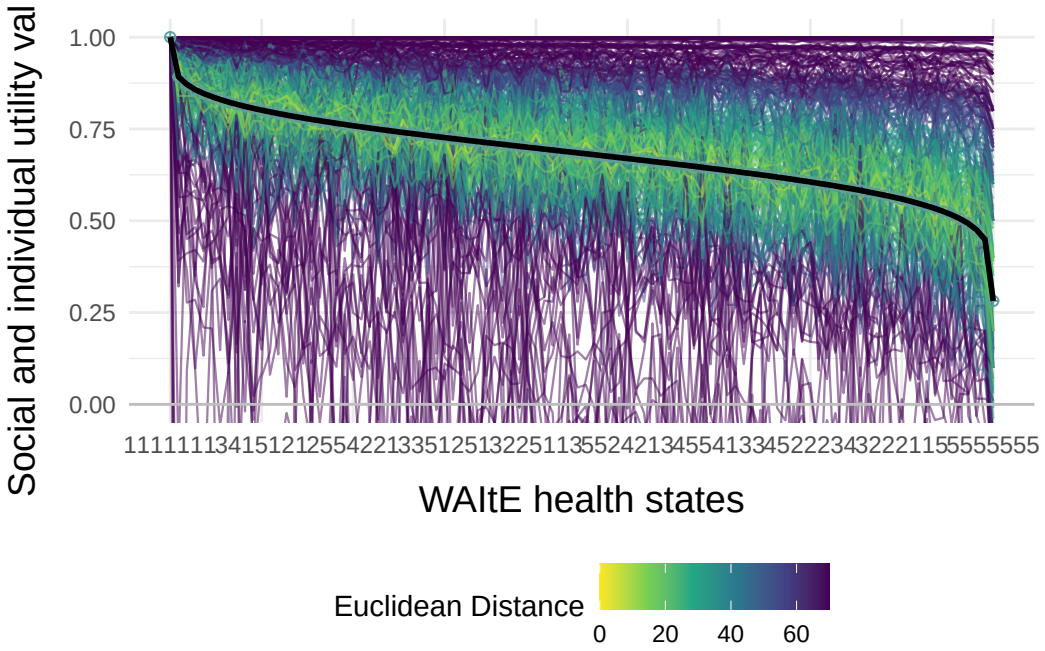

Figure 3: Social and individual utility functions coloured by EUD

PERMANOVA

Table 8 presents the PERMANOVA model results. Presented are within-group sum-of-squares (SS(_W)) for each group individually and for all groups combined, and the corresponding R(^2), pseudo (F), and (p) values. Preference heterogeneity was significantly affected by age ((p) = 0.03), though the amount of variability in preferences that could be explained by age was relatively small (R(^2)=5.7%). Figure 4 presents the difference in preferences between different age groups. Generally, as age increases, health state utility values for each given WAItE health state are higher. That is, younger populations tend to place more disutility on WAItE health problems than older populations. While weight status was not significantly related to preference heterogeneity according to the PERMANOVA model, given the WAItE is a weight-specific measure, it was informative to explore the relationship between preferences and weight status. Though not statistically significant, we can observe a difference in preferences between normal weight and overweight individuals in Figure 5. For a given WAItE health state, overweight individuals in our sample placed less disutility on that state than did normal weight individuals.

Table 8: Results of PERMANOVA – testing for differences in WAItE health state preferences between group characteristics

	Df	SumOfSqs	R2	F	Pr(>F)
age_factor	6	8103.932	0.040	2.012	0.001
weightdata	4	2957.825	0.014	1.102	0.321
educationdata	5	4307.037	0.021	1.283	0.136
occupationdata	7	4180.638	0.020	0.890	0.659
genderdata	3	723.163	0.004	0.359	0.986
ethnicitydata	4	2963.218	0.014	1.104	0.312
Residual	270	181227.225	0.886	NA	NA
Total	299	204463.039	1.000	NA	NA

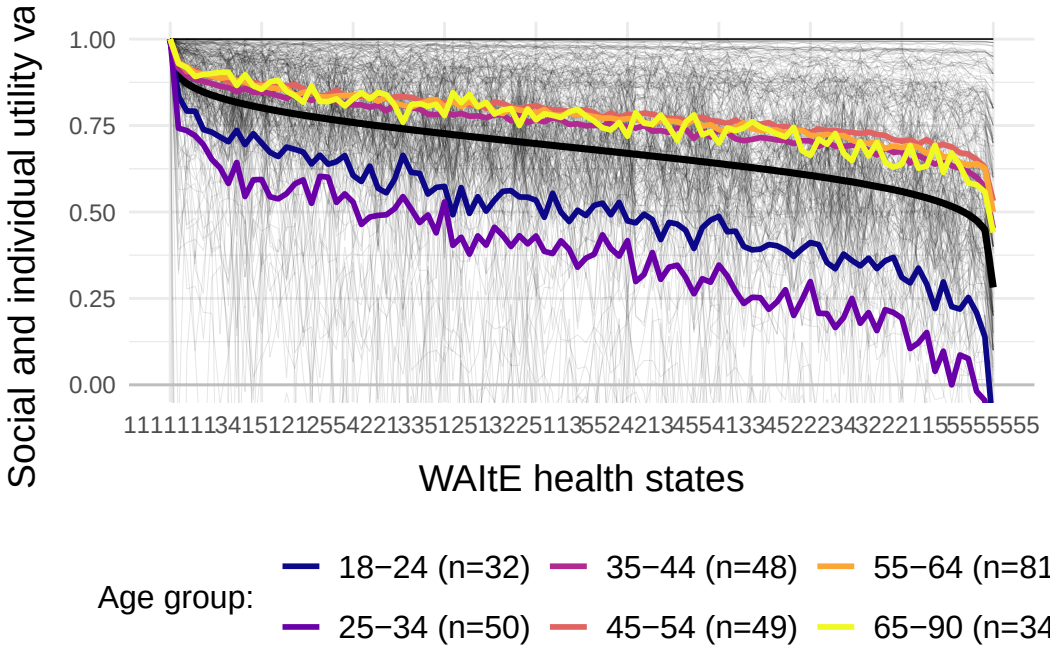

Figure 4: Social and individual utility functions grouped by age status

Sensitivity analysis

EUD2 was estimated for each pairwise comparison of individuals in our study. This yielded a [300 (×) 300] distance matrix with 44,850 unique pairwise comparisons. The mean (SD) and median (IQR) EUD were 34.30 (13.82) and 32.25 (24.54; 41.27). Results from the PERMANOVA2 analysis are presented in Table 9. After exclusion of individual variation in an-

choring responses, weight status and age had a significant impact upon heterogeneity within our sample; though the amount of heterogeneity that was explained by these variables was fairly small (4.9%).

Table 9: Results of PERMANOVA2 – testing for differences in level rating and domain weighting preferences between group characteristics

	Df	SumOfSqs	R2	F	Pr(>F)
age_factor	6	8103.932	0.040	2.012	0.001
weightdata	4	2957.825	0.014	1.102	0.321
educationdata	5	4307.037	0.021	1.283	0.136
occupationdata	7	4180.638	0.020	0.890	0.659
genderdata	3	723.163	0.004	0.359	0.986
ethnicitydata	4	2963.218	0.014	1.104	0.312
Residual	270	181227.225	0.886	NA	NA
Total	299	204463.039	1.000	NA	NA

Discussion

This study is the first time that the OPUF has been used to estimate health state utility values for the WAItE. We obtained a representative sample of high quality data from Prolific, a survey company known for their high quality respondents (Peer et al. 2022). Our average domain weightings and implied ordering were similar to those exhibited in Robinson et al. (XX VIH).

Anchoring of the WAItE PITS state was a difficult procedure that required a number of methodological decisions. We decided to use uncensored responses to the Dead-VAS task which meant that data from one respondent (-99) skewed the mean PITS utility value quite substantially. To mitigate the impact of extreme values on the mean, we conducted winsorization of values lying in the outer 0.1% of the distribution. This practice, while effective at limiting the influence of extreme values on the mean, could understate the genuine variability in the data. Though, it is likely that exclusion of this participant would have had a more detrimental effect to presenting the genuine variability of responses.

The social utility function elicited through this study, and underlying utility value set, present monotonic preferences which behave as we would have expected ex-ante (based on qualitative piloting work). Tiredness and Unhappiness were considered the most important domains while Embarrassment and Sports the least. This finding concurs with qualitative work conducted prior and also is in accordance with previous valuation work done with the WAItE (XX VIH). Prior valuation work, which used a DCE to elicit preferences, yielded latent coefficients which violated the rational choice axiom of monotonicity. In the OPUF, monotonicity is somewhat forced through the choice architecture of the level rating and through the additional prompt

to reconsider responses that are not monotonic. Forced monotonicity, in this context, could be problematic for eliciting unbiased preferences if preferences for certain health states are truly not monotonic. For example, prior qualitative work has suggested that "I almost never get tired" might be preferable to "I never get tired" in some circumstances where respondents are thinking about experiencing insomnia and sleep quality. This being said, the WAItE descriptive system was designed to be a monotonic descriptive system, validated using Rasch analysis, and so having a monotonic utility value set makes logical sense.

Preferences elicited through this study were considerably heterogeneous. This can be understood through the mean EUD value (47.6) but also illustrated in Figure 2 through the deviations of individual PUFs from the social utility function. Following on from prior work (P. Schneider et al. 2024), we estimated EUD by calculating a distance matrix between each pairwise comparison of individual value sets for all 78125 WAItE health states. The implication of estimating distance (preference heterogeneity) by using individual value sets allows for much of the preference heterogeneity that exists to be composed of differences in individual anchoring values (PITS state responses) rather than differences in level ratings and domain weightings. This methodological decision, ultimately, results in the majority of EUD being composed of differences in anchoring values and this finding is important to acknowledge. Anchoring differences are important to present and explore, though in this preference heterogeneity analysis could be drowning out the heterogeneity in level ratings and domain weighting. An example of this can be shown through the age preference heterogeneity in Figure 4. Preference heterogeneity is evident between individuals above and below age 35 and if we consider the mean PITS values for those two subgroups (age (<) 35 = -0.281; age (>) 34 = 0.487) we can see that a clear difference in anchoring responses is evident.

A methodological exploration was conducted as a sensitivity analysis to limit the influence that anchoring variation has on the overall preference heterogeneity. We considered this to be a strength of the research as it offers a new approach to decompose preference heterogeneity into anchoring variation and the difference in level ratings and domain weightings. After exclusion of individual variation in anchoring responses, weight status and age were found to have a significant impact on preference heterogeneity within our sample; though the amount of variation that could be explained was limited. Preference heterogeneity between those of normal weight and those who were overweight is illustrated in Figure Figure 5.

This method of estimating preference heterogeneity should not be considered the gold standard, as only part of the variation in preferences is explored here. It can however be considered an additional option for future researchers that wish to isolate the effect of anchoring responses on overall preference heterogeneity. It is also, to our knowledge, the first time preference heterogeneity has been decomposed in this way with the OPUF.

The value set estimated here offers an alternative choice of preference values to the existing value sets estimated using DCE (shown in Figure **??**). When comparing the anchored coefficients between value sets, one of the key areas of divergence is where levels have been collapsed in the DCE value set. In the OPUF, "I almost never get tired" is given 0.029 compared to 0.064 in the DCE due to collapsing levels. Generally the difference between coefficients that

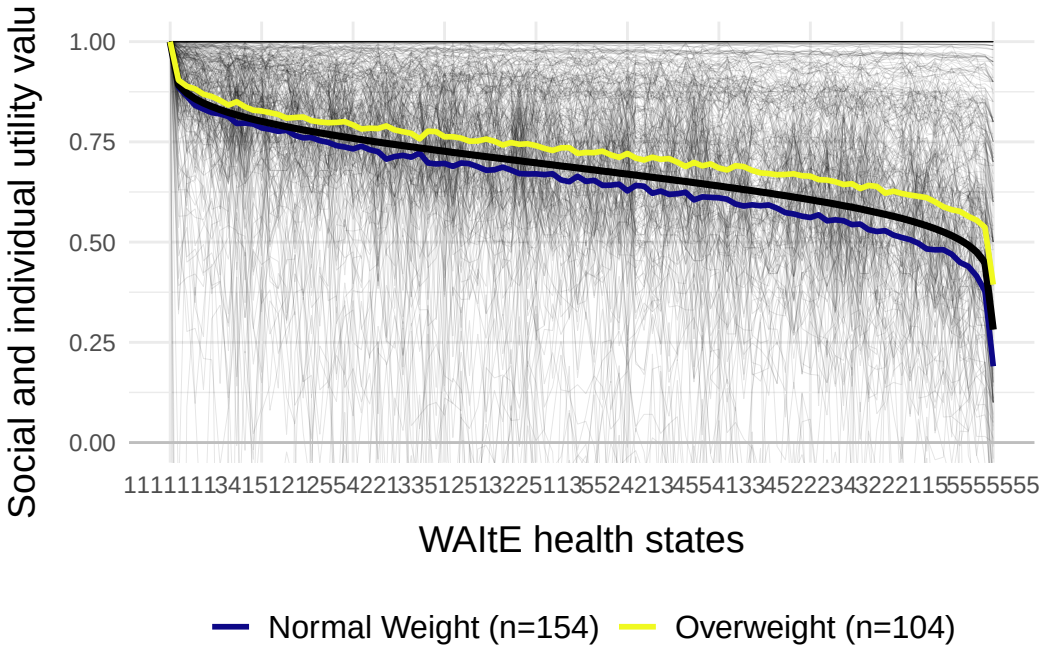

Figure 5: Social and individual utility functions grouped by weight status

have not been 'collapsed' between the value sets is small suggesting that there is comparability to an extent between the value sets. Anchoring values were broadly similar between studies too. The mean PITS utility values between studies were broadly comparable with a maximum range of 0.059. Interestingly, the EQ-VAS anchoring task mean (0.289) was remarkably similar to the OPUF VAS anchoring task mean (0.282) again supporting the use of VAS for elicitation of PITS utility values.

Bibliography

Anderson, Marti J. 2017."Permutational Multivariate Analysis of Variance (PERMANOVA) ." In *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat07841.

Anderson, Marti J., and Daniel C. I. Walsh. 2013. "PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?" *Ecological Monographs* 83 (4). https://doi.org/10.1890/12-2010.1.

Applying Contemporary Statistical Techniques. 2003. https://doi.org/10.1016/b978-0-12-751541-0.x5021-4.

Bana E Costa, Carlos A., and Jean-Claude Vansnick. 1999. "The MACBETH Approach: Basic Ideas, Software, and an Application." In. https://doi.org/10.1007/978-94-017-0647-

6%7B/_%7D9.

Belton, Valerie, and Theodor J. Stewart. 2002. *Multiple Criteria Decision Analysis*. Springer US. https://doi.org/10.1007/978-1-4615-1495-4.

Bray, Nathan, Rhiannon Tudor Edwards, and Paul Schneider. 2024. "Development of a value-based scoring system for the MobQoL-7D: a novel tool for measuring quality-adjusted life years in the context of mobility impairment." *Disability and Rehabilitation*, 1–10. https://doi.org/10.1080/09638288.2023.2297929.

Brodszky, V., S. Plankó, P. Schneider, and N. Devlin. 2023. "PCR108 Pilot Testing the Hungarian Version of Online Elicitation of Personal Utility Functions Tool for Valuing EQ-5D-5L Health States." *Value in Health* 26 (12): S469. https://doi.org/10.1016/j.jval.2023.09.2547.

Danner, Marion, J. Marjan Hummel, Fabian Volz, Jeannette G. Van Manen, Beate Wiegard, Charalabos Markos Dintsios, Hilda Bastian, Andreas Gerber, and Maarten J. Ijzerman. 2011. "Integrating patients' views into health technology assessment: Analytic hierarchy process (AHP) as a method to elicit patient preferences." *International Journal of Technology Assessment in Health Care* 27 (4). https://doi.org/10.1017/S0266462311000523.

Devlin, Nancy J., Koonal K. Shah, Brendan J. Mulhern, Krystallia Pantiri, and Ben van Hout. 2019. "A new method for valuing health: directly eliciting personal utility functions." *European Journal of Health Economics* 20 (2). https://doi.org/10.1007/s10198-018-0993-z.

Hauber, A. Brett, Juan Marcos González, Catharina G. M. Groothuis-Oudshoorn, Thomas Prior, Deborah A. Marshall, Charles Cunningham, Maarten J. IJzerman, and John F. P. Bridges. 2016. "Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force." *Value in Health* 19 (4). https://doi.org/10.1016/j.jval.2016.04.004.

Keeney, R. L., and H. Raiffa. 1979. "Decisions with Multiple Objectives: Preferences and Value Trade-Offs." https://doi.org/10.1109/TSMC.1979.4310245.

Marsh, Kevin, Maarten Ijzerman, Praveen Thokala, Rob Baltussen, Meindert Boysen, Zoltán Kaló, Thomas Lönngren, et al. 2016. "Multiple Criteria Decision Analysis for Health Care Decision Making - Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force." *Value in Health* 19 (2). https://doi.org/10.1016/j.jval.2015.12.016.

Oliveira, Mónica Duarte, Andreia Agostinho, Lara Ferreira, Paulo Nicola, and Carlos Bana E Costa. 2018. "Valuing health states: Is the MACBETH approach useful for valuing EQ-5D-3L health states?" *Health and Quality of Life Outcomes* 16 (1). https://doi.org/10.1186/s12955-018-1056-y.

Peer, Eyal, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. "Data quality of platforms and panels for online behavioral research." *Behavior Research Methods* 54 (4). https://doi.org/10.3758/s13428-021-01694-3.

Robinson, Tomos, and Yemi Oluboyede. 2019. "Estimating CHU-9D Utility Scores from the WAItE: A Mapping Algorithm for Economic Evaluation." *Value in Health* 22 (2). https://doi.org/10.1016/j.jval.2018.09.2839.

Schneider, P P, B van Hout, M Heisen, J Brazier, and N Devlin. 2022. "The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states."

Wellcome Open Res 7: 14. https://doi.org/10.12688/wellcomeopenres.17518.1.

Schneider, Paul, Nancy Devlin, Ben van Hout, and John Brazier. 2024. "Exploring health preference heterogeneity in the UK: Using the online elicitation of personal utility functions approach to construct EQ-5D-5L value functions on societal, group and individual level." *Health Economics (United Kingdom)* 33 (5). https://doi.org/10.1002/hec.4805.

Souza, Allan T., Ester Dias, Ana Nogueira, Joana Campos, João C. Marques, and Irene Martins. 2013. "Population ecology and habitat preferences of juvenile flounder Platichthys flesus (Actinopterygii: Pleuronectidae) in a temperate estuary." *Journal of Sea Research* 79. https://doi.org/10.1016/j.seares.2013.01.005.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple imputation using chained equations: Issues and guidance for practice." *Statistics in Medicine* 30 (4). https://doi.org/10.1002/sim.4067.