
Article:
Schneider, P., Ludwig, K., Marten, O. et al. (2026) Using the OPUF approach to create a value set for the EQ-HWB-S: An exploratory feasibility study [version 2; peer review: 1 approved, 1 approved with reservations]. Wellcome Open Research, 9. 359. ISSN: 2398-502X

https://doi.org/10.12688/wellcomeopenres.21408.2

RESEARCH ARTICLE

REVISED Using the OPUF approach to create a value set for the EQ-HWB-S: An exploratory feasibility study

[version 2; peer review: 1 approved, 1 approved with reservations]

Paul Schneider[1,2], Kristina Ludwig[3], Ole Marten[3], Emily McDool[1], Tessa Peasgood[1], Nancy Devlin[4], Koonal Shah[5], John Brazier[1], Wolfgang Greiner[3], Clara Mukuria[1]

[1]Sheffield Centre for Health and Related Research, The University of Sheffield, Sheffield, England, S1 4DA, UK
[2]Valorem health, Bochum, 44805, Germany
[3]Universitat Bielefeld, Bielefeld, North Rhine-Westphalia, 33615, Germany
[4]School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, 3053, Australia
[5]NICE, London, SW1A 2BU, UK

v2 **First published:** 10 Jul 2024, **9**:359
https://doi.org/10.12688/wellcomeopenres.21408.1
Latest published: 13 Apr 2026, **9**:359
https://doi.org/10.12688/wellcomeopenres.21408.2

Abstract

Objectives

The EQ-HWB-S (EQ Health and Wellbeing Short version) is a new generic measure of health, social care and carer related quality of life, specifically developed to generate utility values. However, its nine dimensions pose a challenge for creating value sets using traditional elicitation techniques, such as time trade-off. A promising alternative method, called Online elicitation of Personal Utility Functions (OPUF) has recently been proposed. The aim of this study was to test the feasibility of using OPUF to create a value set for the EQ-HWB-S.

Methods

We adapted the OPUF tool for the EQ-HWB-S, and piloted it in convenience samples from the UK and Germany. We then conducted an explorative valuation study in both countries in March 2023. We recruited a total of 658 respondents, in four samples: UK (n = 328) and German (n = 110) general population, and German rheumatic disease (n = 110) and diabetes (n = 110) patients. Feasibility was assessed based on completion times, data quality, logical consistency, and respondents' feedback. A demo version of the English OPUF survey is available at: https://valorem.health/eqen-demo Results Median completion times ranged from 11 to 16 minutes across samples. The

Open Peer Review

Approval Status ✓ ?

	1	2
version 2 (revision) 13 Apr 2026		
version 1 10 Jul 2024	✓ view	? view

1. **Maksat Jumamyradov**, University of South Florida, Tampa, USA
Auburn University, Auburn, USA

2. **Ning Yan Gu**, University of San Francisco, San Francisco, USA

Any reports and responses or comments on the article can be found at the end of the article.

OPUF approach yielded well-ordered value set coefficients in all four samples, with precision comparable to a larger EQ-VT study (n = 520). Pain, mobility, and daily activities were the three most important dimensions across all samples. The OPUF-derived UK value set differed from the EQ-VT pilot value set in relative dimension weights, intermediate level positions, and overall scaling.

Conclusion

The OPUF approach was feasible for creating value sets for the EQ-HWB-S in four relatively small samples, including two patient groups. While further validation is needed, OPUF may offer a viable alternative or supplement to traditional valuation techniques for eliciting health state preferences, particularly in settings with limited resources or when eliciting preferences from specific groups.

Plain Language summary

This study looked at a new way to measure people's preferences for different aspects of health, social care and carer related quality of life. The researchers wanted to see if this new method, called OPUF, could be used to create a scoring system for a questionnaire. This scoring system would be used to assign a value to different health and wellbeing states. For example, perfect health and wellbeing would get a higher score than poor health and wellbeing.

In the past, this kind of scoring system has been created by asking people to make choices involving imagining various health and wellbeing problems and how long they were going to live. But this can be difficult for people to do. The OPUF method tries to make this process easier by asking people questions about each aspect of health and wellbeing one at a time.

The researchers tested the OPUF method with people from the UK and Germany, including people from the general public and people with diabetes or rheumatic diseases. The results suggest that OPUF is a promising way to create scoring systems for health and wellbeing questionnaires. It seems to be easier for people to understand than the traditional preference elicitation methods. More research is needed, but this could be a useful tool for improving how we measure health and wellbeing.

Keywords

EQ-HWB-S, OPUF, Preference Elicitation, Health preferences, MCDA, value set, utilities, health-related quality of life, patient preferences

Corresponding author: Paul Schneider (schneider.paulpeter@gmail.com)

Author roles: Schneider P: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Ludwig K**: Conceptualization, Formal Analysis, Methodology, Project Administration, Writing – Review & Editing; **Marten O**: Conceptualization, Formal Analysis, Methodology, Project Administration, Writing – Review & Editing; **McDool E**: Conceptualization, Formal Analysis, Investigation, Writing – Review & Editing; **Peasgood T**: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – Review & Editing; **Devlin N**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Shah K**: Conceptualization, Methodology, Writing – Review & Editing; **Brazier J**: Conceptualization, Methodology, Supervision, Writing – Review & Editing; **Greiner W**: Conceptualization, Supervision, Writing – Review & Editing; **Mukuria C**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Writing – Review & Editing

Competing interests: KL, OM, TP, ND, KS, JB, WG, and CM are members of the EuroQol Group. All Authors have received funding from the EuroQol Research Foundation. PS is a shareholder of a company that develops preference elicitation software. KS is employed by the National Institute for Health and Care Excellence, an organisation that makes evidence-based recommendations informed by utility values. The views expressed in this paper do not necessarily reflect the views of the funder or the authors' institutions. This work was supported by the CINCH Health Economics Research Center, at the University of Duisburg/Essen, Germany.

Grant information: This work was supported by Wellcome [220055; 108903]; the EuroQol Research Foundation [1508-RA].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Schneider P, Ludwig K, Marten O *et al.* **Using the OPUF approach to create a value set for the EQ-HWB-S: An exploratory feasibility study [version 2; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2026, **9**:359 https://doi.org/10.12688/wellcomeopenres.21408.2

First published: 10 Jul 2024, **9**:359 https://doi.org/10.12688/wellcomeopenres.21408.1

REVISED **Amendments from Version 1**

In this revised version, we addressed comments from two reviewers. The abstract was restructured and additional detail was added to the Results section. In the Methods, we clarified that OPUF is a compositional method in which respondents do not make explicit trade-offs between attributed, distinguishing it from decompositional methods such as TTO or DCE. We also expanded the rationale for using different recruitment platforms (Prolific for the UK sample and 'Bilendi' for the German samples) and clarified that all four samples were mutually exclusive. In the Results, we expanded the discussion of differences in completion rates between samples. In the Discussion, we added a more detailed interpretation of the observed differences between the OPUF-derived and EQ-VT pilot value sets and noted that comparisons between value sets created using different methods are conceptually difficult, if not impossible. We also added a sentence on the potential for a hybrid, interviewer-assisted OPUF approach. In the Limitations, we consolidated the discussion of platform and protocol asymmetries. Finally, corrections were made to tables, including a clarification of tied dimension rankings in Table 4 and a missing footnote for education categories in Table 2.

Any further responses from the reviewers can be found at the end of the article

Introduction

The EQ Health and Wellbeing (EQ-HWB) instrument is a new generic measure that goes beyond conventional measures of health-related quality of life (HRQoL) to include carer and social care related quality of life (Brazier *et al.*, 2022; Peasgood *et al.*, 2022). There are two versions: a longer profile measure with 25-items, and a short version, the EQ-HWB-S (EQ Health and Wellbeing Short version), with nine items. The latter was specifically developed for valuation purposes, i.e. to generate utility values on the Quality-Adjusted Life Year (QALY) scale, that is anchored at full health (=1) and dead (=0).

Yet, with nine dimensions, each with five levels, the EQ-HWB-S is still a relatively complex instrument. It can describe a total of (5^9=) 1,953,125 health states, and a simple fixed effects model (commonly used for valuing the EQ-5D-5L) would require estimating a total of (9*4=) 36 coefficients. Although a recent feasibility study has shown that it is feasible to derive a value set for the EQ-HWB-S using time trade-off (TTO) and discrete choice experiments (DCE) (Mukuria *et al.*, 2023), considerable challenges remain:

First, to accurately estimate 36 coefficients, a large number of data points is required – many more data points than usually required for the EQ-5D-5L. As a result, many more participants will need to be recruited and interviewed to create a value set, which may be time consuming and costly. This could potentially limit the extent to which different countries will be able to adopt the EQ-HWB-S to inform policy decision making.

Secondly, in DCE and TTO tasks, participants evaluate entire health states, i.e. they need to consider all nine dimensions simultaneously, when choosing between scenarios or trading-off time. This puts a significant cognitive burden on the participants, who are likely to then use heuristics (Veldwijk *et al.*, 2023). To counter this effect, some overlap between scenarios in a DCE can be used, or each participant can be asked fewer questions. However, these strategies result in a less efficient design, further increasing the sample size required to derive a value set.

An alternative valuation method – called the personal utility function (PUF) approach – was recently developed by Devlin *et al.* (2019). It combines multiple compositional preference elicitation techniques, in three main valuation steps (dimension weighting, level rating, and anchoring), with the aim to derive model coefficients in a more direct and reflective manner. The PUF approach was further refined and adapted for use as a stand-alone online survey, called OPUF, by Schneider *et al.* (2022). It has already been successfully applied in multiple valuation studies for the EQ-5D-5L – including patient and general population samples – as well as for other descriptive systems (Bray *et al.*, 2023).

We hypothesised that (O) PUF could potentially be useful for valuing the EQ-HWB-S, as it allows constructing value sets in relatively small samples. This may enable value elicitation in specific groups (e.g. patients) and/or in shorter time and with fewer resources. The fact that each dimension is considered individually, may also impose a lower cognitive burden on respondents compared with DCE, for example.

The primary objective of this study was to assess the feasibility of using the OPUF method to construct a value set for the EQ-HWB-S in the UK and Germany. We sought to test this with members of the general public as well as with patients. We also compared the constructed value sets between samples, and against an existing pilot value set for the EQ-HWB-S from the UK, created by Mukuria *et al.* (2023) using an adapted EQ-VT protocol (Stolk *et al.*, 2019).

Methods

Overview

This study evaluates the feasibility of using OPUF to create a value set for the EQ-HWB-S. We developed the OPUF survey in two languages, English and German. We first piloted it in qualitative interviews (n = 13) and a quantitative convenience sample of 40 participants. We then conducted valuation studies with respondents from four different samples (UK and German general population, and German patients with diabetes and rheumatic diseases). The feasibility of the method was assessed based on face validity and technical criteria (e.g. completion times, errors, and feedback).

EQ-HWB-S

As mentioned above, the EQ-HWB-S is a generic measure of health, social care and carer related quality of life (Brazier *et al.*, 2022; Peasgood *et al.*, 2022). It consists of nine dimensions, each with five levels taken from different response categories based on difficulty/frequency/severity (see Table 1). For each item, the recall period is seven days. States are defined as combinations of dimension-levels, ranging from the best state, i.e. having no problems ('111111111'), to the worst state, having major problems in all nine dimensions (555555555).

Table 1. EQ-HWB-S descriptive system.

Mobility (MO)
1 No difficulty getting around inside and outside
2 Slight difficulty getting around inside and outside
3 Some difficulty getting around inside and outside
4 A lot of difficulty getting around inside and outside
5 Unable to get around inside and outside
Daily activities (UA)
1 No difficulty doing day to day activities
2 Slight difficulty doing day to day activities
3 Some difficulty doing day to day activities
4 A lot of difficulty doing day to day activities
5 Unable to do day to day activities
Exhaustion (EX)
1 Never exhausted
2 Only occasionally exhausted
3 Sometimes exhausted
4 Often exhausted
5 Exhausted most or all of the time
Loneliness (LO)
1 Never lonely
2 Only occasionally lonely
3 Sometimes lonely
4 Often lonely
5 Lonely most or all of the time
Cognition (CO)
1 Never have trouble concentrating/thinking clearly
2 Only occasionally have trouble concentrating/thinking clearly
3 Sometimes have trouble concentrating/thinking clearly
4 Often have trouble concentrating/thinking clearly
5 Trouble concentrating/thinking clearly most or all of the time

Table 1. *Continued*

Anxiety (AX)
1 Never anxious
2 Only occasionally anxious
3 Sometimes anxious
4 Often anxious
5 Anxious most or all of the time
Sadness/depression (SD)
1 Never sad/depressed
2 Only occasionally sad/depressed
3 Sometimes sad/depressed
4 Often sad/depressed
5 Sad/depressed most or all of the time
Control (CO)
1 Never feel you have no control over your day to day life
2 Only occasionally feel you have no control over your day to day life
3 Sometimes feel you have no control over your day to day life
4 Often feel you have no control over your day to day life
5 Feel you have no control over your day to day life most or all of the time
Physical pain (PA)
1 No physical pain
2 Slight physical pain
3 Moderate physical pain
4 Severe physical pain
5 Very severe physical pain

OPUF – general method

OPUF is a web-based version of the PUF approach, originally developed by Devlin *et al.* (2019). The main difference to traditional valuation methods is that OPUF is based on a different paradigm: it is a compositional, instead of a decompositional, preference elicitation technique (Belton & Stewart, 2002; Keeney *et al.*, 1979; Thokala *et al.*, 2016). While DCE or TTO require participants to evaluate entire health states, from which partial values for dimension-level coefficients are subsequently inferred (= decompositional approach), in the OPUF method, partial values are elicited directly from the participants. Respondents do not make explicit trade-offs between dimensions in a single task. Instead, the swing weighting step requires respondents to consider the relative importance of dimensions, and a value function is then constructed mathematically from the component responses.

OPUF combines multiple decompositional preference elicitation methods, each of which can be modified, depending on the context. The method broadly consists of three steps:

1. Criteria weighting determines the relative importance of the different dimensions of the HRQoL measure, on a scale from 0 to 100.

2. Level rating determines the relative importance of intermediate levels (e.g. mild physical pain) within each dimension, on a scale anchored at the worst (e.g. very severe physical pain = 0) and the best level (e.g. no physical pain = 100).

3. Anchoring maps the values obtained in the previous steps on to the QALY scale, which is anchored at full health (=1) and being dead (=0).

In the following, we briefly outline how the responses from the three steps can be combined to directly derive an additive preference model, containing one coefficient for each dimension-level.

In short, the level ratings are multiplied by the respective dimension weights, and the resulting values are normalised to a 0–1 scale, where 0 represents the best and 1 the worst outcome. The normalised values are then multiplied with the 'anchoring factor'. If the respondent prefers being dead over the worst state, this anchor refers to the position of 'being dead' on a scale between full health and the worst state; otherwise it refers to the position of the worst state on a scale between full health and dead. Since being dead has a utility of 0, we can map the function to the QALY scale. More formally, the model can be defined as follows:

$$c_{ij} = \frac{\left(1 - \frac{l_{ij}}{100}\right) * w_i}{\sum_{k=1}^{n} (w_k) * f(a)}$$

Where c_{ij} refers to the coefficient for *j*th level on dimension *i.* Accordingly, $l_{i,j}$ refers to the rating for level *j* on dimension *i*, w_i refers to the weight for dimension *i*, *n* refers to the number of dimensions, $f(a)$ refers to the anchoring factor. The anchoring factor, $f(a)$ is $f(a) = 1/(1 - v_{pits})$ if $pits \geq dead$, and $f(a) = 1 - v_{dead}$ otherwise. Here, v_{dead} refers to the position of 'being dead' on a 0–1 scale, anchored at full health (=1) and the worst health states (=0); whereas v_{pits} refers to the position of the worst health state on 0–1 scale, anchored at full health (=1) and dead (=0).

To derive the utility for any given EQ-HWB-S state, first the sum of the respective dimension-level coefficients c_{ij} needs to be computed. Since the model is expressed in terms of disutilities, the resulting value needs to be subtracted from 1. Note that, in contrast to DCE and other decompositional methods, OPUF allows constructing this model not only on the group level, but for each individual respondent.

OPUF – online survey implementation and piloting

OPUF has been implemented as a flexible, open source online survey platform, built using modern JavaScript frameworks (Vue.js, Node.js). It consists of modules, such as a 'level rating module', a 'ranking module', etc, which can be combined and customised to accommodate different descriptive systems and types of surveys.

For the present valuation study, we implemented the OPUF method for the EQ-HWB-S in English as well as in German. The initial surveys were based on previous versions used for the EQ-5D-5L (Schneider *et al.*, 2022). The instructions and questions were revised to be more suitable for the EQ-HWB-S. We then piloted the English and German version of the survey in parallel: six and seven one-to-one online interviews were conducted with participants from the UK and Germany respectively. The participants, including members of the general public as well as patients and carers, were recruited from a convenience sample. Members of the research team watched participants complete the survey via 'screen share'. After each task, interviewees were asked debriefing questions, to identify any issues. Changes to one language version were applied to the other, where appropriate. The survey was finally tested without interviewers with a convenience sample of 40 participants, 20 for each language version, recruited via Prolific (Palan & Schitter, 2018) in January 2023.

Below, we describe the final OPUF survey for the EQ-HWB-S in more detail. However, much effort went into optimising the online user experience. The visual presentation of the tasks should be considered an essential aspect of the method. We thus encourage the reader to consult the English and German demo versions of the OPUF survey available at https://valorem.health/eqen-demo and https://valorem.health/eqde-demo, respectively.

Information and informed consent. The survey started with a short introduction to the study, and an informed consent form, which the participants had to read and accept before continuing. Participants were required to click a checkbox to express their consent and the consent information was stored together with the response data in the database.

Warm-up. At the start of the survey, respondents were asked to report their own EQ-HWB-S state, as well as their subjective health and wellbeing, using an adapted version of the EQ-VAS (Rabin *et al.*, 2011), anchored between 'the best' (=100) and 'the worst state you can imagine' (=0). The EQ VAS (visual analogue scale) is not part of the EQ-HWB-S, but it was included to introduce respondents to the concept of rating health and wellbeing on a numeric scale.

Criteria weighting. As the first step, respondents were asked to rank all nine EQ-HWB-S dimensions. They were presented with a list showing the worst levels of each dimension (e.g. Physical pain: very severe) and asked to rank them 'from worst (first) to least bad (last)'. In the next step, the swing weighting task, respondents were asked to assign values between 0 and 100 to swings from the worst to the best level on each dimension (e.g. from very severe to no

physical pain). The dimension ranked first in the previous task was set to a value of 100. This value was fixed and respondents were encouraged to use this as a yard stick to determine the relative value of the other eight dimensions. The order in which the dimensions were displayed was also consistent with the previous ranking.

Level rating. For each dimension, respondents were asked to rate the three intermediate levels (e.g. mild, moderate, severe physical pain) on a scale from 100 (best level) to 0 (worst level). The values for the best and worst level were fixed. If a rating was implausible, e.g. mild pain received a higher, i.e, better rating than moderate pain, a prompt was shown, encouraging the respondent to check their answers again. In the German version of the survey, respondents, who gave and confirmed 3 implausible answers were rejected from the study.

Anchoring. The anchoring task consists of two steps. First, a pairwise comparison between the worst state '555555555' (Scenario A) and 'being dead' (Scenario B). Subsequently, respondents were shown a 0–100 visual analogue scale, where 100 represented 'no health or wellbeing problems', and 0 represented the not-preferred state, i.e. 'being dead', if the respondent preferred the worst state over dead, or the worst health state, if they preferred dead over the worst state. They were asked to indicate where the preferred state lay on that scale. After clicking on the scale, or entering the value into a field, an arrow appeared to indicate the chosen value on the scale. Respondents, who placed '555555555' or 'being dead' at the same value as 'no health or wellbeing problems' had to be excluded from the analysis, because no personal utility function can be constructed for them (all coefficients shrink to zero, or become infinitely large).

Demographic questions and feedback. At the end of the survey, respondents were asked for demographic information (age, sex, education), their experience with illness, and caring responsibilities, and they were also invited to share feedback on the survey.

Construction of personal value functions and sample value sets

For each respondent, a personal EQ-HWB-S value function was constructed using the method described above. It was specified as an additive model with 36 coefficients: four for each of the nine dimensions, representing the utility decrement associated with levels 2–5. Since the implemented anchoring task allows very low utility values (in cases where 'being dead' is placed close to no health and wellbeing problems on the scale), extreme outliers can occur that strongly influence average estimates. To prevent this, the anchor points, i.e. the utility assigned to 555555555, were censored at −1. There may be better methods to deal with this issue, but since this is a feasibility study, the chosen, simple, solution seemed acceptable.

To derive an EQ-HWB-S value set, the personal value functions were aggregated across all respondents within each sample. We report mean coefficient estimates and standard errors.

Sampling

For this valuation study, we recruited participants from four different samples (target sample size n):

1. UK-sample: UK general population (n = 300)
2. GER-sample: German general population (n = 100)
3. DM-sample: German diabetes patients (n = 100)
4. RA-sample: German patients with rheumatic disease (n = 100)

An additional 10% of respondents was recruited as it was expected that a small percentage would be excluded due to reporting impossible anchoring values. All four samples were mutually exclusive; no respondent participated in more than one sample.

The UK sample was recruited through Prolific, a platform for online surveys. The GER, RA, and DM samples were recruited through the marketing panel provider Bilendi, which was chosen because it could provide access to both general population respondents and diagnosed patient groups in Germany. Both general population samples were stratified by age and gender to yield representative samples. For the DM and RA samples, we sought to recruit patients with a spread of demographic characteristics, namely age, sex, and level of education, to ensure a sufficient number of respondents in each stratum. This means, the RA and DM samples are not representative of the respective patient populations.

Feasibility, value set comparisons, and data quality

We assessed the feasibility and practicability of using the OPUF method for creating a value set for the EQ-HWB-S based on multiple criteria.

1) First, we assessed whether the method produces 'plausible' value functions, i.e. are the levels logically ordered? Do the dimension weights have face validity? Could we use the value set to compute QALY values to inform policy decision making?

2) Secondly, we compared the UK value set, constructed in this study using OPUF, with an existing pilot value set created by Mukuria *et al.* (2023), who used the TTO method and an adapted EQ-VT protocol (Stolk *et al.*, 2019). We also compared the OPUF value sets across samples: UK vs GER, GER vs DM, GER vs RA, and RA vs DM. The comparisons were exploratory. We neither pre-specified any hypotheses, nor did we have any definite expectations on how the different value sets should, or should not, differ. Differences between mean coefficient estimates were statistically evaluated using z-tests.

3) Thirdly, we looked at technical feasibility and assessed whether participants were able to complete the survey. This included completion rates, completion times, and the frequency of errors or suspect response patterns.

4) Finally, we considered the feedback provided by the respondents. We assessed how difficult the respondents report that they found it to understand and answer the questions, and how difficult they found it to complete the survey overall. In addition, we asked respondents for feedback in an open-ended question ('Do you have any other comments or feedback?').

Results

Sample

We recruited 328, 110, 110, and 110, for the UK, the GER, the DM, and the RA-sample, respectively. A total of 17 (5%), 9 (8%), 10 (9%), and 11 (10%) respondents were excluded from all further analyses, because they provided invalid responses to the anchoring task (i.e. they set 'being dead' or '555555555' to a value of 100). The survey was conducted within 5 days in March 2023. Sample characteristics are reported in Table 2.

Table 2. Sample characteristics.

	UK	GER	DM	RA
Age				
18–29	62 (19.9%)	18 (17.8%)	5 (5%)	4 (4.0%)
30–39	58 (18.6%)	16 (15.8%)	5 (5%)	8 (8.1%)
40–49	46 (14.8%)	15 (14.9%)	5 (5%)	11 (11.1%)
50–64	102 (32.8%)	27 (26.7%)	14 (14%)	23 (23.2%)
65+	41 (13.2%)	25 (24.8%)	71 (71%)	53 (53.5%)
No response	2 (0.6%)	0 (0%)	0 (0%)	0 (0%)
Sex				
female	161 (51.8%)	49 (48.5%)	43 (43%)	58 (58.6%)
male	149 (47.9%)	51 (50.5%)	57 (57%)	41 (41.4%)
other	1 (0.3%)	1 (1%)	0 (0%)	0 (0%)
Education*				
high	183 (59%)	34 (33.7%)	30 (30%)	29 (29.3%)
medium	86 (28%)	50 (49.5%)	58 (58%)	57 (57.6%)
low	35 (11%)	15 (14.9%)	10 (10%)	13 (13.1%)
N	**311 (100%)**	**101 (100%)**	**100 (100%)**	**99 (100%)**

*Education categories by country. Germany — low: Hauptschule or Realschule without apprenticeship; medium: apprenticeship or Abitur-level qualification without degree; high: Bachelor's degree or above. UK — low: left school at minimum leaving age with no further education; medium: continued education beyond school but no degree; high: Bachelor's degree or above.

The UK and GER samples were broadly representative of the general populations of the UK (in terms of age and sex) and Germany (in terms of age, sex, and education). In the DM-sample, 28 (28%) respondents reported having type I diabetes, 71 (71%) reported having type II, and 1 (1%) reported having another form. In the RA-sample, the five most frequently reported diagnoses were rheumatoid arthritis (62%), arthrosis (41%), chronic pain syndrome (41%), Osteoporosis (11%), and ankylosing spondylitis (10%). More than half (56%) of the respondents reported more than one diagnosis.

EQ-HWB-S OPUF survey results

Warm-up. Only a very small proportion of the respondents (1–5%) reported no health and wellbeing problems in the EQ-HWB-S. Most reported level 2 or 3 problems on at least one dimension. Overall, self-reported health and wellbeing was similar across samples, with the exception of respondents in the RA sample, who consistently reported more severe issues. Details are presented in Table 3 below.

Dimension weights. Table 4 shows the average (SD) dimension weights assigned to the nine EQ-HWB-S dimensions, and the implied ranks. Although physical pain was the most important dimension in all four samples, the average weight was considerably higher in the UK sample than in the other three. Subsequent ranks 2 and 3 were also consistent across samples, and only showed a small difference in the absolute value between the UK and Germany. Ranks 4 to 9 differed across samples. 'Feeling lonely' was the least important dimension in all but the DM sample, in which it was 'feeling anxious'.

Level ratings. The level ratings are shown in Table 5. For all dimensions, average level ratings were well-ordered, i.e. better levels received higher ratings. There are some indications that ratings differed between dimensions, e.g. level 4 physical pain appeared to receive lower ratings than levels 4 in other dimensions. However, the spread around mean estimates was high, and without pre-specified hypotheses, assessments across nine dimensions and four samples do not allow drawing robust inferences.

Table 3. Own EQ-HWB-S state and EQ-VAS.

	UK	GER	DM	RA
No problems [1] – n (%)	16 (5%)	4 (4%)	4 (4%)	1 (1%)
Minor problems [2] – n (%)	106 (34%)	35 (34%)	37 (37%)	20 (20%)
Severe problems [3] – n (%)	106 (34%)	32 (32%)	23 (23%)	48 (48%)
Sum score [4] – median (Q25-Q75)	16 (12–21)	15 (12–19)	14.5 (12–18)	19 (14.5–26)
EQ-VAS – median (Q25-Q75)	74 (61–81)	74 (60–82)	70 (50–80)	59 (40–77.5)

[1]*full health = respondent in state 111111111;*
[2]*respondent reported only level 2 problems (e.g. 21111111, or 212211111);*
[3]*respondent reported level 4 problems on one or more dimensions;*
[4]*sum of dimension levels with a minimum of 9 and a maximum of 45.*

Table 4. Aggregate dimension weights.

	UK		GER		DM		RA	
	rank *	Mean (SD)	rank *	Mean (SD)	rank *	Mean (SD)	rank *	Mean (SD)
MO	2	85.1 (18.8)	2	82.0 (24.0)	1	85.4 (23.1)	2	82.6 (22.7)
UA	3	81.5 (20.0)	3	77.9 (24.0)	3	79.8 (24.7)	3	79.0 (23.4)
EX	5	69.9 (22.8)	6	67.6 (29.3)	7	62.6 (25.4)	6	67.1 (26.7)
LO	9	58.6 (27.4)	9	55.3 (31.2)	8	58.7 (32.0)	9	47.8 (31.6)
CT	8	64.8 (25.1)	7	67.2 (28.6)	5	69.8 (27.9)	5	69.1 (27.2)
AX	7	66.5 (23.9)	8	58.9 (31.0)	9	55.7 (30.0)	8	53.6 (31.9)
SD	4	71.9 (24.9)	5	65.9 (28.8)	6	63.0 (30.5)	7	57.9 (29.7)
CO	6	69.7 (26.4)	4	71.9 (28.7)	4	74.2 (30.4)	4	71.7 (28.7)
PA	1	91.8 (17.3)	1	84.3 (23.9)	1	85.4 (22.8)	1	84.9 (23.4)

MO = Mobility, UA = day-to-day activities, EX = exhausted, LO = lonely, CT = concentrating, AX = anxious, SD = sad/depressed, CO = control, PA = pain;
**rank = implied rank by mean weight. In the DM sample, MO and PA share rank 1 (both mean weight = 85.4).*

Table 5. **Mean (SD) level ratings by sample.**

UK sample	Lvl 2	Lvl 3	Lvl 4
MO	74.2 (22.9)	51.7 (21.3)	23.3 (20.3)
UA	75.1 (21.7)	52.4 (20.3)	23.9 (19.0)
EX	78.9 (20.5)	59.5 (20.0)	30.5 (20.9)
LO	79.2 (21.1)	60.2 (21.5)	34.5 (22.3)
CT	78.8 (20.8)	58.5 (20.9)	32.5 (21.7)
AX	76.7 (22.1)	56.9 (21.2)	29.4 (20.7)
SD	76.3 (22.7)	55.4 (21.4)	26.1 (21.0)
CO	74.6 (23.0)	54.9 (21.7)	28.8 (21.6)
PA	71.5 (23.9)	47.5 (21.3)	18.7 (21.0)
GER sample	**Lvl 2**	**Lvl 3**	**Lvl 4**
MO	71.3 (26.0)	44.7 (24.4)	20.2 (21.2)
UA	74.0 (23.3)	46.9 (24.2)	21.1 (20.0)
EX	76.5 (23.2)	53.6 (24.2)	26.4 (20.4)
LO	73.4 (26.6)	53.0 (26.3)	27.5 (24.3)
CT	73.0 (25.0)	48.9 (23.3)	24.2 (21.0)
AX	73.5 (24.9)	49.6 (26.0)	24.0 (23.6)
SD	73.4 (24.0)	50.3 (25.2)	22.8 (20.9)
CO	70.7 (27.4)	46.9 (25.5)	22.6 (21.9)
PA	75.3 (22.3)	49.1 (24.3)	18.8 (21.0)
DM sample	**Lvl 2**	**Lvl 3**	**Lvl 4**
MO	74.7 (22.0)	49.8 (22.6)	22.1 (18.6)
UA	77.1 (21.6)	50.4 (23.1)	23.9 (21.4)
EX	75.9 (23.3)	55.4 (23.7)	29.3 (23.5)
LO	74.1 (28.0)	56.1 (25.9)	30.7 (25.6)
CT	78.0 (22.0)	55.7 (24.5)	27.6 (25.3)
AX	71.1 (27.4)	51.0 (26.7)	26.8 (22.0)
SD	69.6 (28.1)	47.4 (24.2)	22.9 (20.8)
CO	66.4 (30.1)	45.2 (26.0)	21.2 (22.2)
PA	76.2 (22.6)	51.9 (23.8)	21.0 (21.8)
RA sample	**Lvl 2**	**Lvl 3**	**Lvl 4**
MO	74.9 (23.2)	50.9 (23.1)	25.7 (25.1)
UA	76.1 (22.7)	51.8 (24.2)	25.1 (22.6)
EX	75.1 (24.1)	55.9 (24.3)	24.3 (22.2)
LO	72.8 (28.9)	51.2 (28.7)	26.2 (23.6)
CT	74.6 (23.9)	51.5 (25.5)	24.4 (23.6)
AX	70.9 (27.6)	51.2 (28.2)	26.1 (24.4)
SD	69.3 (28.2)	49.9 (25.5)	23.0 (21.9)
CO	66.4 (29.1)	47.7 (27.4)	24.1 (24.3)
PA	79.3 (22.3)	55.7 (24.5)	21.6 (21.6)

Anchoring. In the UK sample, a total of 96 (31%) respondents preferred the worst EQ-HWB-S state, '555555555', over being dead, whereas in the GER sample, it was 41 (41%), in the DM sample, it was 36 (36%), and in the RA sample, it was 53 (54%), respectively. The median (Q25-Q75) utility value of state '555555555' was −0.07 (−0.36; 0.11) in the UK; −0.03 (−0.68; 0.33) in the GER; −0.08 (−0.10; 0.30) in the DM; and 0.10 (−0.60; 0.54) in the RA sample.

We report medians instead of means, because the anchoring task allowed very negative values. Anchor points below −1 occurred in all samples, with a frequency ranging from 17% in the UK, with a minimum of −216, to 27% in the DM sample, with a minimum of −4999. Note, for subsequent analyses, anchor points were censored at −1.

EQ-HWB-S value sets

We constructed personal, additive value functions for the EQ-HWB-S, consisting of 36 coefficients (anchor points were censored at −1). Although not the focus of this study, we would like to highlight that, in contrast to decompositional methods, such as DCE, OPUF can provide value functions for each individual respondent. To illustrate this point, Figure 1 shows the distribution of preferences underlying the average value set for the RA sample.

Personal value function coefficients were aggregated across respondents to derive value sets for the EQ-HWB-S. Overall, the OPUF approach yielded plausible value sets for all four samples, with consistent, well-ordered coefficients. In this aggregate mean-based model, the worst state, '555555555' had a utility value of −0.15 in the UK, −0.11 in the GER, 0 in the RA, and − 0.17 in the DM sample.

Despite the relatively small sample sizes of 311 to 99 respondents, the confidence intervals around mean coefficients were comparable to the EQ-VT study by Mukuria *et al.* (2023), which was based on TTO and DCE data from 520 respondents, using a hybrid Tobit model with control for heterogeneity.

Figure 2 shows the aggregate mean coefficient estimates and 95% confidence intervals for all four samples. For comparison, the figure also includes the results from Mukuria *et al.* (2023). Table S1 in the Extended data reports detailed results for the aggregate mean coefficient estimates and 95% confidence intervals.

Comparisons between value sets

The comparison between the OPUF and the EQ-VT (Mukuria *et al.*, 2023) pilot value sets for the UK showed a number of notable differences. Although many of the differences in mean coefficients were small in magnitude, about half of them (17 of 36, 47%) were statistically significant. The key drivers of the difference were the following:

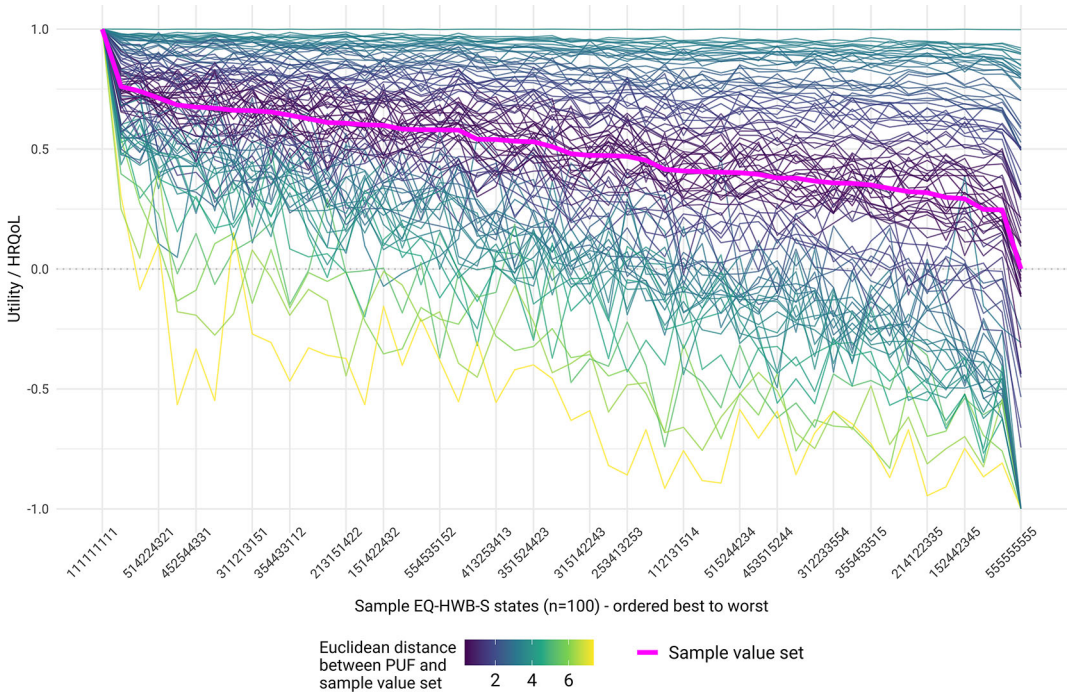

Figure 1. Simplified illustration of the aggregate RA value function (thick line) and the underlying personal value functions of all 99 respondents within the sample. Shown are the utility values for 100 EQ-HWB-S states, ordered best (left) to worst (right). Colours indicate the euclidean distance from the aggregate value set: blue (smaller) to yellow (greater distance).

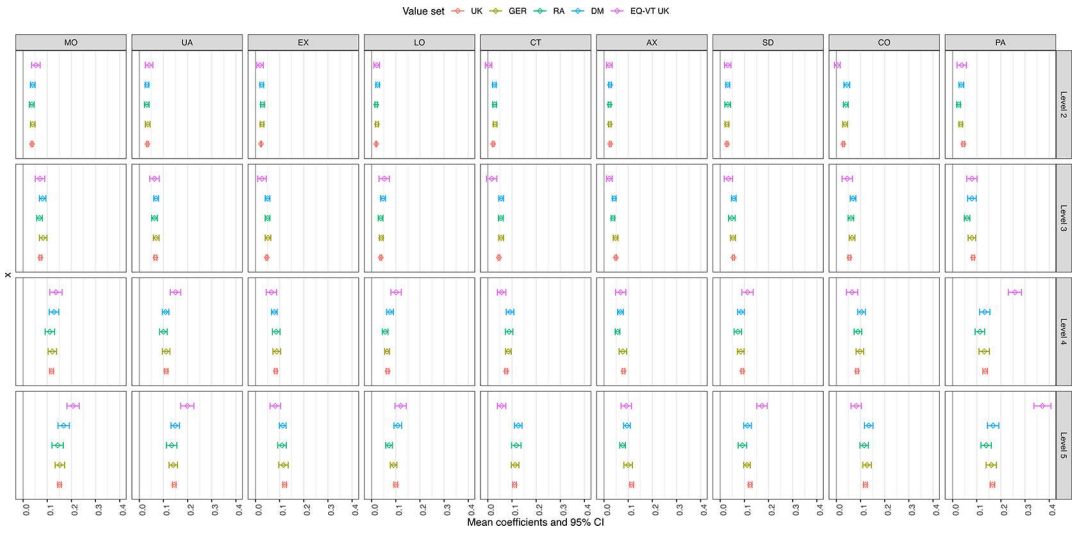

Figure 2. Mean coefficients and 95% confidence intervals for the eight dimensions of the EQ-HWB-S across the four value sets. The dimensions are: MO = Mobility, UA = day-to-day activities, EX = exhausted, LO = lonely, CT = concentrating, AX = anxious, SD = sad/depressed, CO = control, PA = pain. The value sets are: UK = United Kingdom, GER = Germany, RA = Rheumatoid Arthritis, DM = Diabetes Mellitus. The EQ-VT value set for the UK by Mukuria *et al.*, 2023 is also displayed for comparison.

- The relative dimension weights have wider ranges in the EQ-VT model: e.g. the swing from CO5 to CO1 has 15% of the value of the swing from PA5 to PA1. It is thus not surprising that the OPUF UK value set has a larger CO5 coefficient, but a smaller PA5 coefficient.

- While the four most important dimensions are consistent across both OPUF and EQ-VT UK Sample (PA > MO > UA > SD), there was disagreement in the ranking of the other five.

- There are also differences in the relative positions of intermediate levels within dimensions: the EQ-VT model contains logical inconsistencies which required some levels to be combined (AX2 and AX3 = −0.022; CO4 and CO5 = −0.057).

- The lowest utility in the EQ-VT value set is −0.384 (assigned to 555555555), which is lower than the lowest utilities in any of the OPUF value sets (0 to −0.17).

The comparison across the four OPUF value sets showed that coefficients were fairly consistent across samples. Only few statistically significant differences were observed.

All differences in mean coefficients between the OPUF model and the previously created EQ-VT model for the UK (Mukuria *et al.*, 2023), as well as comparisons between the four samples included in this study are reported in Table S2 in the Extended data.

Technical feasibility

Completion rates. For the UK sample, recruited through prolific, the completion rate was about 98%. Only 8 people withdrew after submitting the consent form. For the GER, DM, and RA samples, completion rates were 29%, 24%, and 33% respectively, and thus much lower. Failure to complete the survey could be caused by withdrawal, and in the GER, DM, and RA samples, also by not meeting the eligibility criteria or failing the quality check for illogical level ratings (see below). The lower completion rates in the German samples are likely driven by two (related) factors: (1) differences in panel quality between Prolific and Bilendi; (2) the rejection protocol for illogical level ratings, which was applied in the German samples but not in the UK sample. Additionally, the OPUF survey included complex cursor interactions and drag-and-drop tasks, which should have prevented most automated accounts or bots from completing it. We did not collect data to quantify at what point and why participants discontinued the survey, which limits our ability to disentangle these factors.

Completion times. The median (Q25-Q75) survey completion times were 11.6 (9.0–15.1) minutes in the UK and 11.0 (8.1–14.0) minutes in the GER sample. Both patient samples took a bit longer: 16.1 (11.6–20.4) minutes in the DM, and 15.9 (11.3–22.3) minutes in the RA sample.

Suspect dimension weights. We considered responses suspect if they either included zero dimension weights (indicating that the dimension is not important at all), or included predominantly weights of 100 (indicating a lack of differentiation between dimensions). For the UK, GER, DM, and RA samples, we found that 5 (2%), 2(2%), 5 (5%), and 5 (5%) of the respondents had one or more zero-weights, and that 10 (3%), 9 (9%), 2 (2%), and 4 (4%) assigned a weight of 100 to all nine dimensions.

Illogical level ratings. Table 6 shows the frequencies of illogical level ratings. Ratings within a dimension were considered illogical if a dominated (dominating) level received higher (lower) rating. A notification was shown to the respondents with illogical ratings when they attempted to proceed to the next task. Note that, within the GER, DM, and RA samples, respondents were immediately rejected from the survey, if they ignored the notification and proceeded with illogical level ratings on more than two dimensions.

Impossible anchor points. As noted above, a total of 17 (5%), 9 (8%), 10 (9%), and 11 (10%) respondents in the UK, the GER, the DM, and the RA-sample placed either being dead or state 555555555 at the same position on the VAS as the top anchor 'no health or wellbeing problems'. Since these anchor points do not allow the construction of meaningful value functions, respondents were excluded from any of the analyses reported in this study.

Feedback. Across the four samples, 62–77% of the respondents agreed or strongly agreed with the statement "I found it easy to understand the questions in this survey", and 61–62% agreed or strongly agreed with the statement "I found it easy to answer the questions in this survey". Given that this was a self-administered online survey, we consider the feedback to be positive. For comparison, participants in the interviewer-led EQ-VT study by Mukuria *et al.* (2023), 93% found the task easy to understand, and 62% found it 'easy to decide'. The duration of the OPUF survey was also found to be acceptable. Between 89% and 87% found the length to be 'exactly right'. For more detailed feedback results for each sample, see Table 7.

Across all four samples, a total of 174 (26%) respondents provided feedback as free text. A comprehensive content analysis is beyond the scope of this study, but two common themes immediately emerged from the responses: Firstly, many respondents found the survey 'interesting' or 'thought-provoking' and noted that they had to think about their

Table 6. Number of dimensions with illogical level ratings – n (%).

	UK	GER	DM	RA
No dimension	262 (84%)	81 (80%)	74 (74%)	75 (76%)
1 Dimension	21 (7%)	11 (11%)	17 (17%)	14 (14%)
2 Dimensions	8 (3%)	8 (9%)	9 (9%)	10 (10%)
3+ dimensions *	20 (6%)			

**Participants in the GER. DM, and RA samples were immediately rejected from the survey, if they had illogical level ratings in more than two dimensions.*

Table 7. Quantitative feedback – n (%).

	UK	GER	DM	RA
I found it easy to understand the questions				
Strongly agree	95 (30.5%)	33 (32.7%)	23 (23%)	25 (25.3%)
Agree	143 (46%)	36 (35.6%)	40 (40%)	37 (37.4%)
Neutral	41 (13.2%)	25 (24.8%)	19 (19%)	28 (28.3%)
Disagree	27 (8.7%)	7 (6.9%)	17 (17%)	9 (9.1%)
Strongly disagree	4 (1.3%)	0 (0%)	1 (1%)	0 (0%)
No response	1 (0.3%)	0 (0%)	0 (0%)	0 (0%)

Table 7. *Continued*

	UK	GER	DM	RA
I found it easy to answer the questions				
Strongly agree	50 (16.1%)	21 (20.8%)	23 (23%)	22 (22.2%)
Agree	140 (45%)	41 (40.6%)	39 (39%)	40 (40.4%)
Neutral	59 (19%)	27 (26.7%)	24 (24%)	30 (30.3%)
Disagree	56 (18%)	12 (11.9%)	13 (13%)	7 (7.1%)
Strongly disagree	5 (1.6%)	0 (0%)	1 (1%)	0 (0%)
No response	1 (0.3%)	0 (0%)	0 (0%)	0 (0%)
What do you think about the duration				
Too long	1 (0.3%)	0 (0%)	6 (6%)	0 (0%)
Long	9 (2.9%)	6 (5.9%)	0 (0%)	5 (5.1%)
Exactly right	276 (88.7%)	88 (87.1%)	88 (88%)	86 (86.9%)
Short	21 (6.8%)	5 (5%)	5 (5%)	6 (6.1%)
Too short	2 (0.6%)	2 (2%)	0 (0%)	2 (2%)
No response	2 (0.6%)	0 (0%)	1 (1%)	0 (0%)

answers. Secondly, several comments indicated that some of the questions were 'difficult', 'confusing', 'complex', and/or 'complicated', suggesting potential for further improvement to clarify the instructions and/or scale labels. Only one participant indicated that they experienced technical issues ('Sliders too sensitive').

Discussion

This exploratory feasibility study was the first to investigate the potential of using the OPUF approach for creating value sets for the EQ-HWB-S, and it was one of the first valuation studies for the EQ-HWB-S overall. OPUF yielded plausible, well-ordered value sets in four samples: a UK general population sample, a German general population sample, a sample of people with diabetes mellitus, and a sample of people with rheumatic diseases.

The OPUF method was also found to be technically feasible and well accepted by respondents. Completion rates were high in the UK sample (98%) and moderate (for an online panel also used for marketing surveys) in the other three samples (24–33%). Median survey completion times were 11–16 minutes. In all four samples, we found indications of suspect responses, such as zero-weights or illogical level ratings. However, these were not overly prevalent, and could be reduced further by including additional training exercises or quality checks, at the data collection stage, or applying more stringent exclusion criteria, at the analysis stage. The feedback we received was largely positive, although several respondents commented that some questions were difficult to understand, or confusing, suggesting potential for further improvement. A hybrid approach, with optional interviewer support for respondents who struggle, could be explored in future studies, though the self-administered online format is a deliberate design feature that enables low-cost, scalable data collection.

Despite the relatively small sample sizes of only 100 to 300 respondents per sample, the precision around mean coefficient estimates was comparable to the EQ-VT pilot value set for the UK (Mukuria *et al.*, 2023), which was based on TTO and DCE data from 520 respondents. Our study further validated and confirmed their previous finding that PA, MO, UA, and SD are the four dimensions with the highest relative importance. However, the OPUF value set for the UK also differed in several aspects from the EQ-VT pilot value set, including relative dimension weights, positions of intermediate levels, and overall scaling. These differences are partly expected: the narrower range of dimension weights in the OPUF model directly translates into smaller level 5 coefficients for the most important dimensions, and a less negative worst state utility. The differences in intermediate level positions may reflect the direct level rating approach used in OPUF, compared with the statistical inference from TTO and DCE data in EQ-VT.

The reasons behind the observed differences are not entirely clear. They could be attributed to differences in the selection of respondents, or to the elicitation method. An argument against the former is that we used OPUF to elicit preferences from three other samples (GP, RA, DM), which, even though they differed with respect to demographics and health characteristics, showed high consistency. This finding suggests that the OPUF approach and the EQ-VT protocol may yield different results. This should not be surprising. It is well known that different preference elicitation methods can yield different value sets (Brazier *et al.*, 2017), which immediately raises the question: *which value sets are better?*

In the absence of any gold standard, it seems unclear how to answer this question. Regarding the differences between OPUF and EQ-VT, one can find arguments for either side. On the one hand, OPUF may be more suitable than EQ-VT for creating value sets for the EQ-HWB-S, because it allows respondents to consider each dimension and level individually. This more direct way to elicit preferences may reduce the cognitive burden and improve the quality of the responses. Mukuria *et al.* (2023) observed several inconsistencies in the ordering of coefficients in some of their statistical value set models. This may suggest that asking respondents to evaluate problems in all nine dimensions simultaneously may just be too difficult, and cause respondents to use simplifying heuristics (Veldwijk *et al.*, 2023). On the other hand, as a compositional method, OPUF does not require respondents to make explicit trade-offs between dimensions. TTO and DCE do require such trade-offs, which may help respondents to think more carefully about how they weigh different health and wellbeing problems and their implications. Whether this difference constitutes an advantage or a limitation of OPUF is an open question.

Over and beyond questions related to the *validity* of the different preference elicitation techniques, there are also practical considerations that should be taken into account. OPUF was used as a self-administered online survey. It was easy to set up and implement, and it took most respondents less than 20 minutes to complete. Compared to an interviewer-led TTO or DCE study, considerably fewer resources are required for recruitment, data collection, and analysis. Furthermore, OPUF allowed us to construct well-ordered value sets for the EQ-HWB-S in relatively small samples of patients, and even on the individual level. This may enable researchers to elicit health and wellbeing preferences in specific groups and contexts, in which conventional valuation methods are usually not feasible, such as in clinical trials, or in countries with limited resources.

Limitations

This study was a first explorative feasibility study. It has several limitations and parts of the data analysis are still ongoing. Findings reported in this paper should be considered preliminary.

First, even though we sought to recruit representative samples for the UK and Germany, online panellists are unlikely to be representative of the general population. The RA and DM patient samples further differ in several demographic characteristics from the respective patient populations. Secondly, the use of different recruitment platforms for the UK (Prolific) and German (Bilendi) samples is a limitation, and the rejection protocol for illogical level ratings was only applied in the German samples. Both factors likely contributed to the large difference in completion rates between the UK and German samples, but we did not collect data to quantify their relative contributions. Thirdly, we did not exclude respondents, who had suspicious, potentially implausible responses from the analysis, to not further reduce the sample size. In future versions of the OPUF survey, the prevalence of these types of responses may be reduced by including additional training exercises, quality checks, and/or pre-specifying exclusion criteria. Fourthly, as a compositional preference elicitation method, OPUF has several general limitations. Most importantly, it assumes an additive preference model. It can practically not account for interactions between dimensions. Yet, given the complexity of the EQ-HWB-S instrument, studies using other elicitation methods will probably also not be able to capture interaction effects, due to a lack of statistical power. Finally, although the OPUF method has now been applied in more than 20 individual surveys to create value sets for, among other instruments, the EQ-5D-5L and EQ-5D-Y-3 L, it is still a relatively new approach. The method may require further refinement and/or adaptation for different settings and instruments. More work is needed to validate the approach, show its value in practice, and make it more widely accessible to other researchers.

As immediate next steps, we are investigating the test-retest reliability of the OPUF method using data collected from the GER, DM, and RA samples, about 2 weeks after the first survey was completed. In addition, we are also conducting a comprehensive qualitative study, in order to explore in more detail the experiences of respondents with the OPUF survey, and to gauge their understanding of the questions and responses. Notwithstanding the remaining challenges, we believe that OPUF provides a promising alternative approach to create value sets for the EQ-HWB-S (and other instruments) and that it has the potential to complement conventional valuation methods, such as TTO and DCE, in the future.

Conclusion

We found that the OPUF approach yielded plausible, well-ordered value sets for the EQ-HWB-S. The method was technically feasible and well accepted by respondents. The OPUF value sets for the UK, Germany, diabetes mellitus, and rheumatic diseases samples were fairly consistent, but showed notable differences to the EQ-VT pilot value set for the UK. The OPUF approach may offer an alternative to conventional valuation methods, particularly in settings with limited resources or when eliciting preferences from smaller (patient) groups.

Ethical approval

This study was approved by the Research Ethics Committees of the School of Health and Related Research at the University of Sheffield (ID: 050470) on 24 November 2022, and the University of Bielefeld (ID: 2022–246) on 7 November 2022. All participants provided written informed consent before taking part in the study.

Data availability

Underlying data

The data contains sensitive personal information, including information about health and therefore cannot be made openly available per the decision of the Research Ethics Committees of the School of Health and Related Research at the University of Sheffield. Data will be made available upon reasonable request. To apply for access, interested parties must submit a request to the corresponding author or Clara Mukuria (c.mukuria@sheffield.ac.uk) detailing the purpose of the data request and how the data will be used.

Extended data

Zenodo: Supplementary Material for Schneider *et al.*, Using the OPUF approach to create a value set for the EQ-HWB-S: An exploratory feasibility study. https://doi.org/10.5281/zenodo.11578651 (Schneider, 2024).

This project contains the following extended data:

- Table S1
- Table S2

Acknowledgements

We are grateful to all participants in this study, and to Ernest Law for his insightful discussion of a previous version of this paper at the 2023 EuroQol Plenary meeting in Rome.

This study is mentioned in PS's doctoral thesis (https://etheses.whiterose.ac.uk/33273/), but no methods or results are reported. The paper has not been submitted elsewhere.

References

Belton V, Stewart T: *Multiple criteria decision analysis: an integrated approach.* Springer Science & Business Media; 2002.
Publisher Full Text

Bray N, Tudor Edwards R, Schneider P: **Developing preference-based value sets for the MobQoL-7D: practical application of the Online Elicitation of Personal Utility Functions (OPUF) tool. Working paper.** 2023.

Brazier J, Peasgood T, Mukuria C, *et al.*: **The EQ-HWB: overview of the development of a measure of health and wellbeing and key results.** *Value Health.* 2022; **25**(4): 482–491.
PubMed Abstract | **Publisher Full Text**

Brazier J, Ratcliffe J, Saloman J, *et al.*: *Measuring and valuing health benefits for economic evaluation.* Oxford: Oxford University Press; 2017.
Reference Source

Devlin NJ, Shah KK, Mulhern BJ, *et al.*: **A new method for valuing health: directly eliciting personal utility functions.** *Eur J Health Econ.* 2019; **20**(2): 257–270.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Keeney RL, Raiffa H, Rajala DW: **Decisions with multiple objectives: preferences and value trade-offs.** *IEEE Trans Syst Man Cybern.* 1979; **9**(7): 403.
Publisher Full Text

Mukuria C, Peasgood T, McDool E, *et al.*: **Valuing the EQ Health and Wellbeing Short using time trade-off and a Discrete Choice Experiment: a feasibility study.** *Value Health.* 2023; **26**(7): 1073–1084.
PubMed Abstract | **Publisher Full Text**

Palan S, Schitter C: **Prolific.ac—a subject pool for online experiments.** *J Behav Exp Finance.* 2018; **17**: 22–27.
Publisher Full Text

Peasgood T, Mukuria C, Brazier J, *et al.*: **Developing a new generic health and wellbeing measure: psychometric survey results for the EQ-HWB.** *Value Health.* 2022; **25**(4): 525–533.
PubMed Abstract | **Publisher Full Text**

Rabin R, Oemar M, Oppe M, *et al.*: *EQ-5D-5L user guide. Basic information on how to use the EQ-5D-5L instrument.* Rotterdam: EuroQol Group; 2011, 22.

Schneider P: Supplementary material for Schneider *et al.*, using the OPUF approach to create a value set for the EQ-HWB-S: an exploratory feasibility study. 2024. In: Wellcome Open Research. *Zenodo.* [Dataset]. 2024.
Publisher Full Text

Schneider PP, van Hout B, Heisen M, *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 Res.* 2022; **7**: 14.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Stolk E, Ludwig K, Rand K, *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): 23–30.
PubMed Abstract | **Publisher Full Text**

Thokala P, Devlin N, Marsh K, *et al.*: **Multiple criteria decision analysis for health care decision making--an introduction: report 1 of the ISPOR MCDA Emerging Good Practices Task Force.** *Value Health.* 2016; **19**(1): 1–13.
PubMed Abstract | **Publisher Full Text**

Veldwijk J, Marceta SM, Swait JD, *et al.*: **Taking the shortcut: simplifying heuristics in Discrete Choice Experiments.** *Patient.* 2023; **16**(4): 301–315.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ?

Version 1

Reviewer Report 10 October 2024

https://doi.org/10.21956/wellcomeopenres.23676.r94903

?

Ning Yan Gu
[1] University of San Francisco, San Francisco, California, USA
[2] University of San Francisco, San Francisco, California, USA

- Can Prolific platform recruit German general population as well to match n=300 for the UK general population? Also, since UK sample had high completion rate of 98%, compared with relatively lower completions in other samples where Prolific was not used, it may suggest sample/cohort selection bias.
- Please explain if these samples mutually exclusive; for example, those who included in disease population are not included in the general population.
- In the abstract, please know the conclusions were included in results section, need to move to conclusion and need to add more data/info. to the results before draw conclusions.
- Please know Table 7 showcased that there are still a large proportion of the respondents did not find the survey to be easy; hence, maybe interviewer assistant could be needed.
- Please explain clearly whether or not or how the critical element of 'trade-off' was sustained / reserved in the process of asking respondents weighing each item individually.
- Understand this is preliminary work; since it is preliminary early-stage work, is it feasible to compare utilities derived by using OPUF with utilities derived by using traditional EQ-VT? If there no significant difference, then the benefit of minimize respondents' cognitive burden is truly great and verified. If there are differences or major differences, need to understand and explain those differences.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Health Economics and Outcomes Research, QALY, elicitations, behavioral analyses, etc.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 09 August 2024

https://doi.org/10.21956/wellcomeopenres.23676.r90530

Maksat Jumamyradov
[1] University of South Florida, Tampa, Florida, USA
[2] Health Outcomes Research and Policy, Auburn University, Auburn, Alabama, USA
[3] University of South Florida, Tampa, Florida, USA
[4] Health Outcomes Research and Policy, Auburn University, Auburn, Alabama, USA

The paper tests the feasibility of Online elicitation of Personal utility Functions (OPUF) approach in constructing a value set for the EQ-HWB-S. Authors demonstrate the feasibility of this innovative approach in creating the value set with small samples, which seems to be a promising alternative to the traditional valuation techniques. The paper is well-written, easy to follow and presents a method that will be a valuable addition to the literature.

I only have a few minor comments:

1. I tried to complete the demo version (English) of OPUF survey through the link provided in the manuscript, however, after a certain point, the “Continue” button stopped working and I

wasn't able to finish the survey. Maybe it happens because it is a demo version. I was just curious to see how the questions are formulated.

2. Table 4, column for DM has two 1st rank dimensions, MO and PA.

3. Just a suggestion for future studies, in the second step of the "Criteria weighting" in the OPUF demo, it might be helpful for the respondents if you put the attributes in the same order as they were selected in the first step. For instance, if I select PA (rank 1), MO (rank 2), UA (rank 3) etc. in the first step of criteria weighting, it would be helpful if all 9 attributes appear in the same order in the second step of the criteria weighting when I need to score them between 0 and 100. Otherwise, it will be difficult for me to remember in which order I selected the attributes, which may lead to "illogical level ratings".

4. Table 2 has a star next to "Education", but I wasn't able to find the explanation of what the star means.

5. Subsection illogical level ratings: Could you please explain why the GER, DM and RA samples were immediately rejected from the survey if they ignored illogical level ratings notification, but the UK sample wasn't? Also, would this be a potential explanation why you have such big difference in completion rates (i.e. 98% UK sample vs ~30% German sample). I don't think bots alone would cause such a difference in completion rates.

Thank you very much for the opportunity to review this paper.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Health preference research, discrete choice experiments, econometrics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.