Test-retest reliability of the Online elicitation of Personal Utility Functions (OPUF) approach for valuing the EQ-HWB-S

Aisha Moolla[a], Paul Schneider[a], Ole Martin[b], Clara Mukuria[a], Tessa Peasgood[a]

[a]Sheffield Centre for Health and Related Research, University of Sheffield, Sheffield, UK

[b]Department of Health Economics and Health Care Management, Bielefeld University, Bielefeld, Germany

Abstract

Introduction: The EQ Health and Wellbeing Short (EQ-HWB-S) is a new instrument developed to generate utility values. However, the complexity of the instrument makes traditional preference elicitation techniques challenging to apply. The Online elicitation of Personal Utility Functions (OPUF) approach has recently been tested as an alternative that seems to overcome some of these challenges. The aim of this study was to evaluate the test-retest reliability of the OPUF approach for valuing the EQ-HWB-S.

Methods: The OPUF EQ-HWB-S survey was completed twice by 220 German participants, including a general population sample (73) and a sample of patients with diabetes or rheumatic disease (147), two weeks apart. The test-retest reliability of the outcomes was assessed at an individual and aggregate level. Each component of the survey, including dimension rankings and swing weights, level weights, and anchoring factors were assessed for reliability. Continuous data were compared using the intraclass correlation coefficient (ICC), and ranking data were compared using Spearman's correlation coefficient. Individual and aggregate level utility decrements were compared using the ICC and t-tests.

Results: In the analysis of the dimensions, 36% of participants had significantly correlated dimension ranks, with 42% of participants choosing the same top ranked dimension. ICC values for individual dimension swing weights were consistently below 0.59, with 70% of ICC values indicating poor agreement. For individual level weights, ICC values showed poor agreement in 70% and moderate agreement in 30% of responses. In the analysis of the individual pairwise comparison task, the unweighted kappa was 0.64 (95% CI: 0.54-0.75) showing moderate agreement; however, the ICC comparing individual-level anchoring factors was 0.12 ($p<0.05$), indicating poor agreement. The t-test indicated that the means of aggregate utility decrements were similar for all dimensions.

Conclusion: Our findings suggest that the OPUF approach produces reliable value sets for the EQ-HWB-S on the aggregate group level. However, further refinement may be needed to improve consistency on the individual level.

INTRODUCTION

The EQ Health and Wellbeing (EQ-HWB) is a new instrument designed to measure health, social care, and carer-related quality of life (1). The short version, the EQ-HWB-S, which has nine dimensions with five response levels, was specifically designed to derive utility values essential for economic evaluations of health and social care interventions. Dimensions include mobility (MO), daily activities (DA), exhaustion (EX), loneliness (LO), cognition (CG), anxiety (AX), sadness/depression (SD), control (CO), and physical pain (PA).
EQ-HWB utility values have been estimated using time trade-off (TTO) and discrete choice experiment (DCE) tasks (2) which are conventional decompositional preference elicitation techniques. However, the length and complexity of the EQ-HWB poses a challenge for these techniques. The EQ-HWB describes an extensive 1,953,125 health states, a tiny proportion of which are used in the estimation of the 36 utility decrements comprising the utility-algorithm. Conventional techniques necessitate large sample sizes to generate these decrements, and TTO and DCE tasks may impose cognitive burdens on participants, especially when contemplating nine dimensions simultaneously (1–3). An alternative that can address these limitations is a compositional preference elicitation technique, the personal utility function (PUF) approach, that allows for the direct elicitation of partial values (4,5). The PUF approach allows estimation of utility functions for individuals and at the aggregate level and due to the approach, could be used to generate a utility function using a very small sample (n=1) (4).

An online PUF (OPUF) approach for EQ-HWB was recently tested in a UK and German population. The UK sample consisted of a general population sample and the German sample consisted of a general population sample and two patient samples (diabetes and rheumatic disease). Both tests produced plausible value sets and showed that it is feasible to use this technique (unpublished). However, the reliability of the OPUF has not been assessed to date. Given the increasing uptake of the OPUF in eliciting utility values (5–7), it is crucial to determine whether the technique produces consistent results. This study aimed to evaluate the test-retest reliability of the OPUF method in valuing the EQ-HWB-S in the German sample. Details on the UK sample will be reported elsewhere (unpublished).

MATERIALS AND METHODS

Sample

Adult participants were recruited from Germany using a market research company's online panel for participation in a validation study (unpublished). The initial test was completed between 6 and 11 March 2023, and the retest was completed between 20 and 30 March 2023. Data were collected from a total of 330 participants. Of these participants, 110 were representative of the general population in terms of age and gender (which we refer to as the GP-sample), 110 were patients with diabetes (DM-sample), and 110 were patients with rheumatic disease (RA-sample). The same 330 participants were then invited to complete the survey again after 2 weeks. The participant preferences towards different health and wellbeing states were expected to remain consistent during this timeframe.

Preference elicitation survey

The OPUF employs a three-step valuation process to derive utility decrements for each dimension-level (5). The first step aims to obtain swings weights for each dimension based on its relative importance. The next step aims to generate the level weights for each intermediate level of each dimension, which are anchored at the worst and best level. The final step aims to generate an anchoring factor that maps all health states on the quality-adjusted life year (QALY) scale, which is anchored at full health (100) or the dead state (0). These valuation steps are broken down and information is generated via a survey.

The EQ-HWB-S OPUF survey was delivered through an open-source online survey platform. The survey included questions related to the EQ-HWB-S and demographic questions (https://valorem.health/eqen-demo). The survey structured was as follows:

1. An introduction to the study and informed consent
2. A warmup task in which participants reported their own EQ-HWB state and an adapted version of the EQ-VAS
3. A dimension ranking task in which each EQ-HWB dimension is ranked 'from worst (first) to least bad (last)' in a list format.
4. The dimension swing weights (between 0 and 100) were then elicited. The participant provided swing weights by indicating the value assigned to moving from the worst to the

best level in each dimension. Moving from worst to the best level in the top ranked dimension from the previous task was given a fixed score of 100 and improvements in the other eight dimensions were scored relative to this improvement.

5. Intermediate levels of each dimension were assigned a weight (between 0 and 100) by asking participants to rate the intermediate levels. The best and worst levels were anchored at 100 and 0, respectively. When more severe levels were assigned higher weights than less severe levels, the response was considered illogical.
6. A pairwise comparison between the worst state ('555555555') and dead state was performed to elicit which scenario the participant preferred.
7. Anchoring was performed by assigning a value to the preferred state from the previous task. This was done by having participants indicate where the preferred state lies on a scale from 0 (representing the less preferred state in the previous task) and 100 (representing no health or wellbeing problems). Anchoring values were censored at -1.
8. Finally, participants provided demographic data and overall feedback.

Data analysis

Estimating the utility decrements

Figure 1. Survey tasks with corresponding outputs used to derive utility decrements

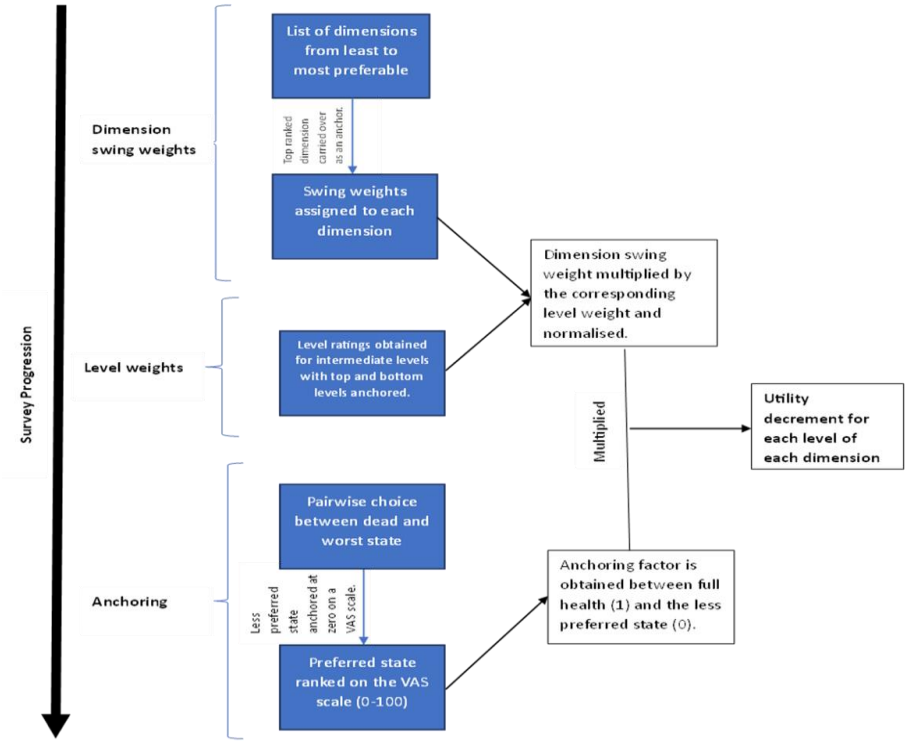


Figure 1. shows that components of the tasks and the process of deriving utility decrements. Utility decrements were estimated by multiplying the level rating by the corresponding dimension weight, the product of which was then normalised between 0 (best) and 1 (worst). An anchoring factor was estimated based on the position of the dead or worst health state (depending on the choice made by the participant) on a scale between full health and the worst state or dead. The normalised value was multiplied by the anchoring factor to produce a utility decrement, which was subtracted from 1 to produce a utility value. An additive model was used to derive the utility value of each health state for each participant as well as the entire population.

Test-retest

Test-retest reliability was assessed at both the aggregate and respondent level for the total sample population (GP + RA + DM samples), the general population (GP) sample, and the combined patient sample (RA + DM samples). Illogical responses were identified when participants selected zero-weights or weights of 100 for dimension swing weights or when severe levels were rated lower than less severe levels. Participants were excluded when they had more than two illogical responses. They were also excluded if they indicated indifference between full health and either death or state 555555555 in the anchoring task as meaningful value functions could not be derived from this.

General respondent characteristics, such as demographics, completion times, and the number of illogical responses, were reviewed for both test and retest samples.

A. Dimension ranks

At the aggregate level, consistency was assessed by examining the proportion of participants who gave the same dimension the top ranking in both tests, reported as percentage agreement. The top ranked dimension acts as an anchor in the subsequent question, making this selection vital to the overall ranking of dimensions. A percentage agreement of ≥ 70% was considered adequate agreement between test and retest (8).

For individual-level dimension rankings, Spearman's rank correlation coefficient was calculated for each participant as the sample size was too small to construct accurate confidence intervals around the weighted kappa statistic (9). The size of the correlation was interpreted based on the following rho thresholds: negligible = 0.00-0.30, low = 0.30-0.50,

moderate = 0.50−0.70, high = 0.70-0.90, and very high = 0.90-1.00 (10). The proportion of participants who had significant positive correlations with a rho greater than 0.30 were reported.

B. Dimension swing weights

For individual-level dimension weights, we assessed reliability using the two-way mixed effects intraclass correlation coefficient (ICC) as both the degree of correlation and the agreement are relevant. A mixed effects model was selected as these tests (the test and retest) were the only "raters" of interest. There was no need to generalise inferences to other tests (11). The ICC strength of agreement was classified as follows: poor = < 0.40, moderate = 0.41–0.59, good = 0.60–0.74, and excellent > 0.75 (12).

C. Level ratings

The reliability of intermediate level ratings was compared between the test and retest (the top and bottom levels are used as anchoring points and set to 0 and 100, respectively). Individual-level rating weights were then compared using the ICC as described above. Respondents with any illogical level ratings (e.g., more severe levels were rated better than less severe levels) were excluded from this part of the analysis as participants with illogical responses were thought to have interpreted the question incorrectly and the test aimed to evaluate consistency rather than understanding.

D. Anchoring factor

In the first step of the anchoring task, participants were asked to choose between the worst state '555555555' (Scenario A) and 'being dead' (Scenario B). The consistency of this task was compared using percentage agreement. The agreement in this test was also assessed using the unweighted kappa statistic. The following cutoffs were used: poor = 0-0.2, fair = 0.21-0.40, moderate = 0.41–0.60, strong = 0.61–0.80, and near complete = > 0.81 (13). In order to include all participants regardless of their selection (Scenario A or the Dead state) the ICC was calculated for the anchoring factor rather than the score produced using the visual analogue scale. The ICC was estimated for the overall group and for those who prefer death or the worst state separately. Utility values were censored at -1 for this analysis.

E. Utility decrements and value set

Both individual- and aggregate-level utility decrements were assessed for reliability between the test and retest. The value set, based on aggregate utility decrements, for both the test and retest were produced. The aggregate means of the decrements were compared using a paired t-test. However, the significance of the t-test is driven by the between-individual variances of the test and retest values, which may lead to an inaccurate result in the case of high variances (14). As such, the empirical distributions of the aggregate level decrements were also compared using a Q-Q plot to perform a visual comparison and a two-sample Kolmogorov-Smirnov test to ascertain statistical significance. This test compares the cumulative distribution functions of each test and is based on both the location and shape of the distributions (15,16). Analysis of the distribution is also critical to understand due to its incorporation into economic analyses.

Individual-level decrements were compared using the ICC. Individual-level rankings of all health states scored based on the individual's utility decrements were also compared using Spearman's rank correlation decrements.

Impact of other factors on test-retest

Linear regression was performed to assess whether age, gender, sample group, change in self-reported health, or change in speed of completion predicted the cumulative difference in individual-level utility decrement values between test and retest. The cumulative difference was calculated as the sum of the absolute differences across all utility decrements. This metric was selected due to the very small values attached to the differences in utility decrements.

In all tests, p-values were considered significant at <0.05. Normally distributed data were presented as the mean and standard deviation (SD) and non-normally distributed data were presented as the median and interquartile range (IQR). All statistical analyses were carried out using R version 4.3.1. This study was approved by the University of Bielefeld (ID: 2022-246). Informed consent was obtained for all participants.

RESULTS

Sample characteristics

In total, 330 and 257 participants completed the initial and retest surveys respectively. There were 110 participants in each sample (GP, DM, and RA) (Table 1). After the exclusion of participants, with illogical responses (n=21 and 18, respectively) or unusable responses because as they could not be matched to test responses (n=19), the final analysis sample was 220 (66.67%) participants. Table 1 describes the demographic and test characteristics of all participants included in the analysis.

Reported gender demographics remained consistent, while education level and age varied between the test and retest. In the total sample, 4 participants reported a different age category and 24 reported a different education level. Completion rates and self-reported health when using a visual analogue scale (VAS) remained similar in all groups.

Table 1. Sample characteristics for each sub-group in the test and retest

	Test			**Retest**		
	Total population	**GP sample**	**Patient sample**	**Total population**	**GP sample**	**Patient sample**
Total sample	330	110	220	257	85	172
Excluded	21	0	21	18	5	13
Total included	309	110	199	239	80	159
Unmatched responses				19	7	12
Matched sample				220	73	147
Age n (%)						
18-29	17 (7.7%)	12 (16.4%)	5 (3.4%)	17 (7.7%)	12 (16.4%)	5 (3.4%)
30-39	22 (10%)	12 (16.4%)	10 (6.8%)	22 (10%)[#]	12 (16.4%)	10 (6.8%)[#]
40-49	20 (9.1%)	11 (15.1%)	9 (6.1%)	21 (9.6%)[#]	11 (15.1%)	10 (6.8%)[#]
50-64	48 (21.8%)	22 (30.1%)	26 (17.7%)	48 (21.8%)[#]	22 (30.1%)	26 (17.7%)[#]
65+	113 (51.4%)	16 (21.9%)	97 (66.0%)	112 (51.0%)[#]	16 (21.9%)	96 (65.3%)[#]
Gender n (%)						
Female	101 (45.9%)	31 (42.5%)	70 (47.6%)	101 (45.9%)	31 (42.5%)	70 (47.6%)
Male	118 (53.6%)	41 (56.2%)	77 (52.4%)	118 (53.6%)	41 (56.2%)	77 (52.4%)
Other	1 (0.5%)	1 (1.4%)	0 (0%)	1 (0.5%)	1 (1.4%)	0 (0%)
Education n (%)						
High	63 (28.6%)	26 (35.6%)	37 (25.2%)	66 (30.0%)[#]	27 (37.0%)[#]	39 (26.5%)[#]
Medium	132 (60.0%)	38 (52.1%)	94 (63.9%)	120 (54.5%)[#]	33 (45.2%)[#]	87 (59.2%)[#]
Low	22 (10.0%)	8 (11.0%)	14 (9.5%)	30 (13.6%)[#]	11 (15.1%)[#]	19 (12.9%)[#]
Not indicated	3 (1.4%)	1 (1.4%)	2 (1.4%)	4 (1.8%)[#]	2 (2.7%)[#]	2 (1.4%)

Self-reported health						
VAS score median (IQR)	70 (50-80)	75 (60-85)	69 (41-80)	70 (50-81)	80 (65-86)	69 (41-80)
Survey completion times						
Completion time (min) median (IQR)	14.2 (10.4-20.0)	11.1 (8-16.1)	15.8 (11.6-22.1)	13.0 (9.4-19.0)	11.4 (7.5-17.3)[#]	13.5 (10.4-20.0)

[#]indicates differences in test and retest

Dimension ranks and swing weights

At the aggregate level, all samples had low consistency between test and retest for the top-ranked dimension. In the total sample, 93 of the 220 (42.27%) participants chose the same top ranked dimension. In the GP and patient samples, 36/73 and 57/147 participants, respectively, had consistently chosen the same top ranked dimension. This resulted in a percentage agreement of 49.32% and 38.78%, which was considered low. Similarly, consistency in individual-level dimension rankings was low, with only 36.46% of participants in the total sample, 41.10% of participants in the GP sample, and 34.01% of participants in the patient sample having significant positive correlations between tests.

For the swing weights, the individual-level ICC for five dimensions (MO, EX, LO, CO, and PA) were classified as having poor agreement; while four were considered moderate (DA, CG, AX, and SD) in the total sample (Table 2). In the GP sample, the ICC strength of agreement was classified as poor for three dimensions (MO, AX, and CO); moderate for five dimensions (DA, EX, LO, SD, and PA); and good for one dimension (CG). In the patient sample, there was poor agreement in all except one (AX) dimension, which had moderate agreement.

Table 2. ICC values comparing dimension weights by sample

	Total population	GP-sample	Patient sample
Mobility	0.25**	0.39**	0.18*
Daily activities	0.42**	0.54**	0.37**
Exhaustion	0.37**	0.46**	0.33**
Loneliness	0.37**	0.45**	0.32**
Cognition	0.46**	0.61**	0.39**
Anxiety	0.41**	0.36**	0.43**
Sadness/depression	0.47**	0.59**	0.40**
Control	0.34**	0.34**	0.33**
Pain	0.38**	0.45**	0.34**

*p<0.05, **p<0.01

Agreement level: poor = < 0.40, moderate = 0.41–0.59, good = 0.60–0.74, and excellent > 0.75

Level ratings

In the total population sample, 45 (20.45%) and 46 (20.90%) of participants in the test and retest, respectively, produced illogical responses and were excluded for this part of the analysis. In total, 152 participants were included in the analysis, with 23 participants consistently producing illogical responses in both tests. ICC values analysing level weights were statistically significant (Table 3). These values showed poor agreement in 70.37% and moderate agreement in 29.63% of the level ratings.

In the GP sample, 12 (16.44%) and 13 (17.81%) participants were excluded in the test and retest, respectively, due to illogical responses. In total, 54 participants were included in the final analysis, with 6 participants consistently producing illogical responses. All except one (UA level 4) ICC value, were statistically significant. Of those that were significant, more than half (54%) showed moderate agreement and the rest (46%) showed poor agreement.

In the patient sample, 33 (22.45%) participants were excluded in both the test and retest, respectively, due to illogical responses. In total, 114 participants were included in the final analysis, with 17 participants consistently producing illogical responses. Two ICC values were not statistically significant and showed no agreement between test and retest values. Of those that were significant, most (77%) showed poor agreement and the rest (23%) showed moderate agreement.

Table 3. ICC values comparing absolute intermediate levels by sample

	Total population		
	Level 2	**Level 3**	**Level 4**
Mobility	0.17*	0.26**	0.23**
Daily activities	0.40**	0.39**	0.16*
Exhaustion	0.34**	0.40**	0.40**
Loneliness	0.27**	0.33**	0.35**
Cognition	0.41**	0.38**	0.41**
Anxiety	0.28**	0.41**	0.41**
Sadness/depression	0.35**	0.49**	0.33**
Control	0.27**	0.44**	0.45**
Pain	0.34**	0.34**	0.40**
	GP-sample		
	Level 2	**Level 3**	**Level 4**
Mobility	0.23*	0.38**	0.26*
Daily activities	0.38**	0.42**	0.15
Exhaustion	0.35**	0.39**	0.40**

Loneliness	0.29**	0.30*	0.29*
Cognition	0.44**	0.46**	0.48**
Anxiety	0.38**	0.42**	0.43**
Sadness/depression	0.45**	0.49**	0.50**
Control	0.46**	0.50**	0.57**
Pain	0.44**	0.31*	0.45**
	Patient sample		
	Level 2	**Level 3**	**Level 4**
Mobility	0.13	0.19*	0.22*
Daily activities	0.41**	0.38**	0.16
Exhaustion	0.34**	0.41**	0.40**
Loneliness	0.26**	0.34**	0.40**
Cognition	0.39**	0.35**	0.38**
Anxiety	0.24**	0.41**	0.39**
Sadness/depression	0.31**	0.48**	0.21*
Control	0.18*	0.42**	0.40**
Pain	0.28*	0.35*	0.37**

*p<0.05, **p<0.01

Agreement level: no agreement = 0, poor = < 0.40, moderate = 0.41–0.59, good = 0.60–0.74, and excellent > 0.75

Anchoring

In the total sample, the percentage agreement for the pairwise comparison task was 82.73%, indicating a high level of consistency. The unweighted kappa also showed good agreement, with a value of 0.64 (95% CI: 0.54-0.75). Overall, 117 participants consistently preferred the dead state, and 69 participants consistently preferred the worst health state. Only 34 participants changed their responses between the test and retest. In the total sample, the mean anchoring factors in the test and retest were -0.09 and -0.14, respectively. The overall ICC when comparing anchoring factors was 0.12 (confidence interval: -0.015-0.25), indicating poor agreement. When considering only those who consistently selected the dead state or the worst health state, the ICC was 0.12 (confidence interval: -0.057-0.30) and 0.12 (confidence interval: -0.12-0.34), indicating poor agreement.

The percentage agreement in the GP and patient samples for the pairwise comparison task was 83.56% and 82.31%, respectively, which was considered good agreement. Similarly, the unweighted kappas were 0.65 (95% CI: 0.48-0.83) and 0.64 (95% CI: 0.52-0.76), indicating good agreement. In the GP sample, 39 participants consistently prefer the dead state, and 21 participants consistently prefer the worst health state, with 13 participants changing their responses between the test and retest. In the patient sample, 73 participants

consistently prefer the dead state, and 48 participants consistently prefer the worst health state, with 26 participants changing their responses.

The mean anchoring factors were -0.13 and -0.08 in the test for the GP and patient samples, respectively. The mean retest anchoring factors were -0.14 and -0.14 in the GP and patient samples, respectively. The ICC produced when comparing anchoring factors in the overall group was -0.00066 ($p>0.05$) and 0.16 ($p<0.05$), indicating no agreement and poor agreement in the GP and patient samples. Among those in the GP sample who consistently chose the dead state, the ICC was -0.017 ($p>0.05$), and among those who consistently chose the worst health state, the ICC was 0.57 ($p<0.01$). This indicates that those who select the worst state produce a more consistent anchoring value than those who prefer the dead state. Among those in the patient sample who consistently chose the dead state, the ICC was 0.19 ($p>0.05$), and among those who consistently chose the worst health state, the ICC was -0.040 ($p>0.05$). This indicates no agreement in either group.

Utility decrements and value set

Table 4 shows the ICC values produced when comparing the 36 individual level utility decrements. In the overall sample, 35 of the 36 ICC values were significant. Of those, the ICC showed poor agreement in 23 decrements and moderate agreement in 12 decrements.

In the GP sample, the ICC values were significant in 33 of the decrements. Of those that were significant, 1 ICC value indicated good agreement (CG2), 21 indicated moderate agreement, and 11 indicated poor agreement. In the patient sample, the ICC values were significant in 32 of the decrements. Of those that were significant, 8 indicated moderate agreement and 24 indicated poor agreement.

Table 4. ICC values when comparing individual level utility decrements

	Total population	GP sample	Patient sample
Mobility level 2	0.13*	0.13	0.13
Mobility level 3	0.36**	0.42**	0.32**
Mobility level 4	0.47**	0.50**	0.45**
Mobility level 5	0.49**	0.55**	0.47**
Daily activities level 2	0.23**	0.24*	0.23**
Daily activities level 3	0.30**	0.42**	0.24**
Daily activities level 4	0.42**	0.47**	0.38**
Daily activities level 5	0.51**	0.51**	0.51**

Exhaustion level 2	0.25**	0.12	0.35**
Exhaustion level 3	0.36**	0.47**	0.30**
Exhaustion level 4	0.45**	0.56**	0.39**
Exhaustion level 5	0.38**	0.50**	0.32**
Loneliness level 2	0.10	0.09	0.11
Loneliness level 3	0.21**	0.23*	0.20**
Loneliness level 4	0.43**	0.55**	0.39**
Loneliness level 5	0.35**	0.55**	0.34**
Cognition level 2	0.25**	0.67**	0.10
Cognition level 3	0.29**	0.53**	0.17*
Cognition level 4	0.42**	0.53**	0.38**
Cognition level 5	0.50**	0.52**	0.48**
Anxiety level 2	0.17**	0.22*	0.13
Anxiety level 3	0.27**	0.27*	0.25**
Anxiety level 4	0.34**	0.29**	0.38**
Anxiety level 5	0.37**	0.35**	0.38**
Sadness/depression level 2	0.30**	0.55**	0.19**
Sadness/depression level 3	0.34**	0.49**	0.26**
Sadness/depression level 4	0.42**	0.41**	0.43**
Sadness/depression level 5	0.43**	0.46**	0.42**
Control level 2	0.21**	0.31**	0.17*
Control level 3	0.21**	0.24*	0.19*
Control level 4	0.38**	0.42**	0.35**
Control level 5	0.42**	0.41**	0.43**
Pain level 2	0.31**	0.42**	0.27**
Pain level 3	0.30**	0.31**	0.29**
Pain level 4	0.34**	0.25*	0.40**
Pain level 5	0.41**	0.35**	0.44**

*p<0.05, **p<0.01

Agreement level: no agreement = 0, poor = < 0.40, moderate = 0.41–0.59, good = 0.60–0.74, and excellent > 0.75

The aggregate level utility decrements were also compared between the test and retest. The mean overall utility decrement was similar (0.08) in both the test and retest. Figure 2 shows the small absolute differences in aggregate level decrements between the test and retest in each dimension for the total sample. The mean absolute difference was 0.004. Figure S1 provides a graphical representation of the distributions of the utility decrements. The Q-Q plots (Figure S1) show that the distributions of the aggregate utility decrements are similar between the test and retest, with many plots appearing to intercept at zero and lie on the 45 degree line.

Figure 2. Aggregate level utility decrements for levels three and five in the test and retest (total sample)

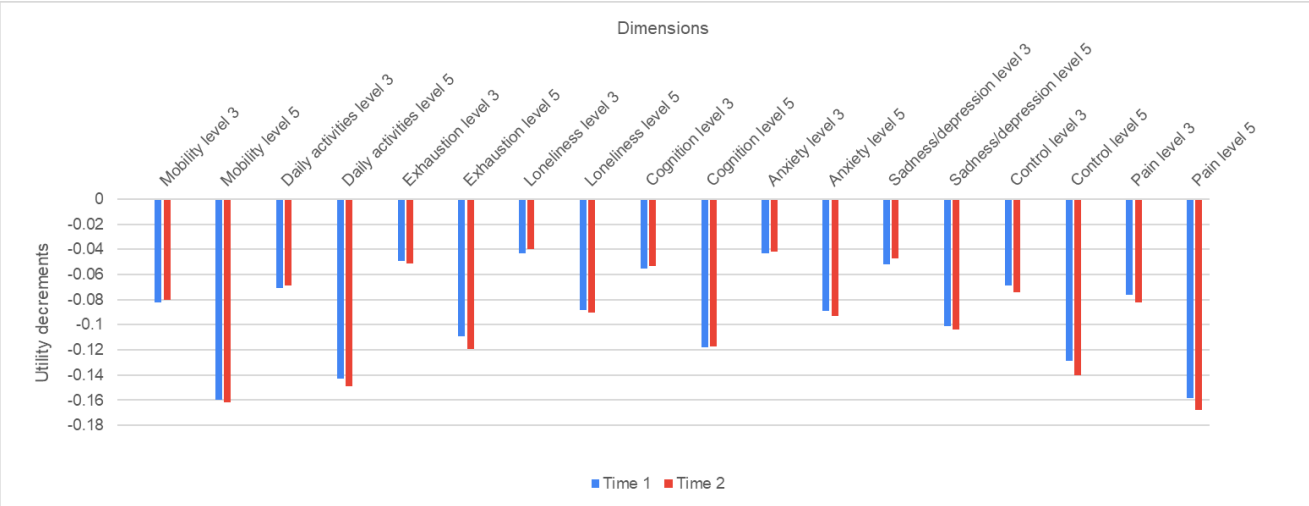


Table 5 shows the Kolmogorov-Smirnov test and a paired t-test results from the comparison of aggregate level utility decrements. In the total, GP, and patient samples, the D and t statistics are not statistically significant with the exception of D-statistic for EX5 in the total sample and PA3 in the patient sample, indicating that test and retest distribution of these decrements were not significantly different. The t-statistic was not statistically significant for SD3 in the GP sample, indicating differences in means. This may not be evident in the KS test given that this test is partly driven by the distribution, giving the mean less influence over the overall significance of the test.

The final health state rankings were compared between the test and retest using Spearman's rank correlation test. The rho was 0.26 ($p<0.05$), 0.26 ($p<0.05$), and 0.26 ($p<0.05$) in the total, GP, and patient samples, respectively, indicating a low to negligible positive monotonic relationship between the test and retest health state ranks.

Table 5. Kolmogorov-Smirnov and t-test results by sample

	Total population		GP sample		Patient sample	
	D_n (p-value)	t (p-value)	D_n (p-value)	t (p-value)	D_n (p-value)	t (p-value)
MO2	0.06 (0.84)	-0.63 (0.52)	0.10 (0.89)	0.69 (0.49)	0.06 (0.95)	-1.15 (0.25)
MO3	0.07 (0.60)	0.35 (0.73)	0.12 (0.64)	0.96 (0.34)	0.06 (0.95)	-0.23 (0.82)
MO4	0.05 (0.95)	0.54 (0.59)	0.07 (0.20)	-0.002 (1.00)	0.05 (1.00)	0.66 (0.51)
MO5	0.08 (0.53)	-0.31 (0.76)	0.11 (0.78)	0.004 (1.00)	0.07 (0.89)	-0.36 (0.72)
DA2	0.08 (0.45)	0.67 (0.50)	0.12 (0.64)	1.19 (0.24)	0.09 (0.61)	-0.02 (0.98)
DA3	0.05 (0.95)	0.45 (0.65)	0.08 (0.97)	0.94 (0.35)	0.07 (0.80)	-0.07 (0.94)
DA4	0.08 (0.45)	-0.85 (0.39)	0.11 (0.78)	0.25 (0.80)	0.11 (0.35)	-1.24 (0.22)
DA5	0.08 (0.53)	-0.86 (0.39)	0.10 (0.89)	0.29 (0.77)	0.12 (0.28)	-1.27 (0.21)
EX2	0.07 (0.61)	-1.06 (0.29)	0.11 (0.78)	-1.08 (0.28)	0.10 (0.52)	-0.41 (0.69)
EX3	0.10 (0.18)	-0.64 (0.52)	0.15 (0.38)	-0.83 (0.41)	0.09 (0.61)	-0.22 (0.82)
EX4	0.05 (0.95)	-0.49 (0.62)	0.14 (0.50)	-0.13 (0.89)	0.07 (0.81)	-0.49 (0.62)
EX5	0.13 (0.04)*	-1.75 (0.08)	0.18 (0.20)	-1.34 (0.19)	0.13 (0.17)	-1.25 (0.21)
LO2	0.06 (0.84)	0.67 (0.50)	0.10 (0.89)	0.42 (0.67)	0.06 (0.95)	0.52 (0.61)
LO3	0.05 (0.95)	0.67 (0.50)	0.11 (0.78)	0.32 (0.75)	0.06 (0.95)	0.59 (0.56)
LO4	0.05 (0.95)	0.31 (0.76)	0.12 (0.64)	0.06 (0.96)	0.07 (0.89)	0.32 (0.75)
LO5	0.08 (0.45)	-0.44 (0.66)	0.11 (0.78)	-0.77 (0.44)	0.09 (0.61)	0.03 (0.97)
CG2	0.07 (0.69)	-0.34 (0.74)	0.12 (0.64)	0.14 (0.89)	0.07 (0.80)	-0.46 (0.65)
CG3	0.06 (0.84)	0.64 (0.52)	0.10 (0.89)	0.71 (0.48)	0.04 (1.00)	0.35 (0.73)
CG4	0.07 (0.69)	0.69 (0.49)	0.15 (0.38)	-0.02 (0.98)	0.08 (0.71)	0.80 (0.43)
CG5	0.08 (0.45)	0.27 (0.78)	0.14 (0.50)	-0.28 (0.78)	0.07 (0.80)	0.48 (0.63)
AX2	0.07 (0.69)	0.20 (0.85)	0.10 (0.89)	-0.30 (0.77)	0.10 (0.43)	0.49 (0.63)
AX3	0.06 (0.84)	0.16 (0.88)	0.08 (0.97)	0.71 (0.48)	0.05 (0.98)	-0.45 (0.65)
AX4	0.05 (0.98)	0.38 (0.70)	0.11 (0.78)	1.42 (0.16)	0.05 (0.98)	-0.90 (0.37)
AX5	0.10 (0.27)	-0.69 (0.49)	0.08 (0.97)	0.87 (0.39)	0.11 (0.35)	-1.67 (0.10)
SD2	0.06 (0.76)	0.74 (0.46)	0.10 (0.89)	1.92 (0.06)	0.07 (0.89)	-0.15 (0.88)
SD3	0.06 (0.76)	1.46 (0.15)	0.11 (0.78)	2.23 (0.03)*	0.05 (1.00)	0.31 (0.76)
SD4	0.08 (0.53)	1.02 (0.31)	0.15 (0.38)	1.23 (0.22)	0.09 (0.61)	0.35 (0.73)
SD5	0.09 (0.38)	-0.63 (0.53)	0.10 (0.89)	0.56 (0.58)	0.12 (0.28)	-1.14 (0.25)
CO2	0.10 (0.84)	0.21 (0.84)	0.14 (0.50)	-0.31 (0.76)	0.05 (0.98)	0.44 (0.66)
CO3	0.07 (0.61)	-0.84 (0.40)	0.11 (0.78)	-0.55 (0.58)	0.08 (0.71)	-0.63 (0.53)
CO4	0.09 (0.38)	-0.97 (0.33)	0.12 (0.64)	-1.02 (0.31)	0.09 (0.61)	-0.44 (0.66)
CO5	0.12 (0.07)	-1.74 (0.08)	0.12 (0.64)	-0.58 (0.56)	0.13 (0.17)	-1.74 (0.08)
PA2	0.1 (0.22)	-1.18 (0.24)	0.10 (0.89)	0.43 (0.67)	0.13 (0.17)	-1.55 (0.12)
PA3	0.13 (0.06)	-0.91 (0.37)	0.11 (0.78)	-0.16 (0.88)	0.16 (0.04)*	-0.98 (0.33)
PA4	0.04 (0.99)	-0.39 (0.70)	0.10 (0.89)	-0.19 (0.84)	0.07 (0.80)	-0.35 (0.73)
PA5	0.08 (0.53)	-1.08 (0.28)	0.12 (0.64)	-0.16 (0.87)	0.11 (0.35)	-1.28 (0.20)

*p<0.05, statistically significant

Dimensions: MO, Mobility; DA, Daily activities; EX, Exhaustion; LO, Loneliness; CG, Cognition; AX, Anxiety; SD, Sadness/depression; CO, Control; PA, Pain

Regression analysis

Table 6 shows the output of the linear regression to assess whether sample group, change in self-reported health on the VAS scale, or difference in completion times could predict the cumulative difference in utility decrements values. An interaction term was added between age and sample group due to multicollinearity between these variables. The cumulative difference was higher for males, those with longer completion times, and those aged 50-65 years when in the RA sample.

Table 6. Regression analysis results

Coefficient	Estimate	P-value
Age: 30-39 years[#]	0.18	0.13
Age: 40-49 years[#]	0.08	0.51
Age: 50-64 years[#]	0.13	0.22
Age: 65+ years[#]	0.10	0.41
Sample: DM[¶]	-0.05	0.80
Sample: RA[¶]	-0.13	0.56
Gender: male[†]	-0.10	0.01[*]
Change in self-reported health	0.00037	0.78
Difference in completion times	0.0042	0.01[*]
Age: 30-39 years[#] and sample: DM[¶]	-0.24	0.33
Age: 40-49 years[#] and sample: DM[¶]	-0.14	0.59
Age: 50-64 years[#] and sample: DM[¶]	0.47	0.04[*]
Age: 65+ years[#] and sample: DM[¶]	0.05	0.82
Age: 30-39 years[#] and sample: RA[¶]	-0.16	0.56
Age: 40-49 years[#] and sample: RA[¶]	0.04	0.89
Age: 50-64 years[#] and sample: RA[¶]	0.04	0.88
Age: 65+ years[#] and sample: RA[¶]	0.04	0.86

*p<0.05

[#]Reference group: 18-29 years, [¶]Reference group: GP sample, [†]Reference group: Female

DISCUSSION

The findings of this study highlight several noteworthy aspects regarding the test-retest reliability of the OPUF EQ-HWB-S. There was a lack of consistency demonstrated in the separate tasks at the individual and aggregate level where applicable with a notable exception of the pairwise task in the anchoring step. This had an impact on the agreement of the utility values at the individual level which were not consistent, but this lack of consistency was not shown in the aggregate utility values.

When considering dimension rankings, only 42.27% of total participants chose the same top ranked dimension. This suggests a notable degree of variability in individual responses. The ICC values assessing dimension swing weights across all samples revealed predominantly moderate to poor agreement, indicating a lack of consistency in the ranking of health dimensions. While one dimension in the GP sample showed good agreement, the overall pattern suggests that participants struggled to maintain consistent responses over the test-retest period. The inconsistency observed in intermediate level weights further emphasises the challenges associated with individual-level responses, with ICC values consistently indicating moderate to poor agreement.

Interestingly, the kappa values (0.6) and percentage-agreement values (83%) derived from pairwise comparisons demonstrated good agreement across all samples, contradicting the inconsistency observed in preceding tasks. This is similar to previous studies in Chinese and German participants using DCE tasks with additional dimensions for the SF-6Dv2 and QLU-C10D instruments. In these studies, the kappa was 0.528 and 0.605, and the percentage agreement was 76.4% and 80.2%, respectively, for the DCE task (17,18). This suggests that comparisons might be the easiest task to complete consistently for participants. This discrepancy with poor results in preceding tasks may suggest that while participants were able to provide relatively consistent pairwise comparisons, they encountered difficulties when asked to rank dimensions or provide weights individually. This could be indicative of the cognitive processes involved in comparing and ranking health states, and further investigation into the reasons behind this inconsistency is warranted.

The poor agreement observed in ICC values for anchoring factors underscores potential challenges in maintaining consistent reference points across test and retest sessions. This

finding may be particularly relevant in understanding the impact of variations in participants' comprehension or interpretation of the survey instructions. This does, however, differ from previous valuation studies using a VAS scale administered by interviewers. In a Spanish population, the mean ICC value when comparing EQ-5D health state valuation using VAS was 0.90, indicating high agreement (19). Similarly, in a British population, the mean ICC was 0.78 when using the VAS scale (20).

At the individual level, utility decrements demonstrated varying degrees of agreement, mostly moderate to poor, with the sole exception of one decrement in the GP sample displaying higher consistency. Conversely, aggregate-level utility decrements exhibited a notably high degree of similarity between the test and retest sessions. The consistency observed at the aggregate level contrasts with the overall lack of uniformity in health state rankings at the individual level. Notably, the narrow range across which health state utility values exist contributed to this inconsistency, as even slight alterations in utility values could result in significant shifts in rankings.

The digital nature of the survey may contribute to the challenges faced by participants. The regression results indicate that being male and spending more time may explain some of the differences. Time is often used to indicate quality respondents in online surveys, but it can also be an indicator of a lack of understanding or poor engagement where respondents start surveys and take time longer because they know they may be excluded if they are too fast. The steep learning curve and the attention-demanding nature of the tasks may have led to the high degree of inconsistency (21). These issues may be likely to occur in those with poor digital literacy, suggesting the potential need to include interviewers (22).

The identified challenges in individual-level responses emphasise the need for further qualitative research to explore which specific tasks participants find challenging and the underlying reasons for inconsistencies. Insights gained from qualitative studies can inform refinements in the survey design, potentially enhancing participant understanding and reducing response variability. Once design adjustments are made, based on qualitative findings, reassessment of the test-retest reliability of the instrument is essential. This iterative process of refinement and re-evaluation is crucial for ensuring the validity and reliability of health-related quality of life assessments.

Participant understanding and engagement appear to be the most crucial aspects for improving results of the OPUF EQ-HWB-S. When tasks are clear and easier to comprehend, such as DCE rather than TTO tasks, consistency appears to improve. This is evident when we compare the current results to the high consistency observed in another German sample where participants completed an online survey using DCE tasks (18). Conversely, when complex valuation tasks, such as Person Trade Off tasks, are used, ICC values reduce to between -0.17 and 0.82 (23).

The observed differences between the test and retest for individual and aggregate-level utility decrements raise questions about the applicability of the OPUF approach with a sample size of one (4). While aggregate-level utility decrements demonstrated high similarity between test and retest, individual-level utility decrements exhibited poor to moderate agreement, indicating potential limitations in the survey's ability to capture stable "personal" utility functions. Inconsistencies in individual-level tasks, which are intended to be simpler than other elicitation tasks, bring into question the validity of the final utility values. Further research is required to explore these concerns further. Additionally, an analysis of the minimum sample size required for the meaningful application of the OPUF approach is warranted. Understanding the trade-off between individual and aggregate-level reliability is vital for researchers and policymakers seeking to implement this approach in diverse contexts.

CONCLUSION

In conclusion, while the OPUF EQ-HWB-S holds promise as a tool for assessing health-related quality of life, this study illustrates that the OPUF approach produces reliable value sets for the EQ-HWB-S on the aggregate group level only. Individual level tasks still lack reliability when using this approach. This necessitates careful consideration and refinement of the OPUF method in order to produce consistent individual-level responses.

REFERENCES

1. Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, et al. The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results. Value in Health. 2022 Apr 1;25(4):482–91.

2. Mukuria C, Peasgood T, McDool E, Norman R, Rowen D, Brazier J. Valuing the EQ Health and Wellbeing Short Using Time Trade-Off and a Discrete Choice Experiment: A Feasibility Study. Value in Health. 2023 Jul 1;26(7):1073–84.

3. Veldwijk J, Marceta SM, Swait JD, Lipman SA, de Bekker-Grob EW. Taking the Shortcut: Simplifying Heuristics in Discrete Choice Experiments. Patient. 2023 Jul 1;16(4):301–15.

4. Devlin NJ, Shah KK, Mulhern BJ, Pantiri K, van Hout B. A new method for valuing health: directly eliciting personal utility functions. Eur J Health Econ. 2019 Mar;20(2):257–70.

5. Schneider PP, van Hout B, Heisen M, Brazier J, Devlin N. The Online Elicitation of Personal Utility Functions (OPUF) tool: a new method for valuing health states. Wellcome Open Res. 2022;7:14.

6. Schneider P, Blankart K, Brazier J, van Hout B, Devlin N. Using the Online Elicitation of Personal Utility Functions Approach to Derive a Patient-Based 5-Level Version of EQ-5D Value Set: A Study in 122 Patients With Rheumatic Diseases From Germany. Value Health. 2023 Dec 27;S1098-3015(23)06242-3.

7. Bray N, Tudor Edwards R, Schneider P. 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. Disabil Rehabil. 2024 Jan 11;1–10.

8. Kazdin AE. Artifact, bias, and complexity of assessment: The ABCs of reliability. Journal of Applied Behavior Analysis. 1977;10(1):141–50.

9. Fleiss JL, Cicchetti DV. Inference About Weighted Kappa in the Non-Null Case. Applied Psychological Measurement. 1978 Jan 1;2(1):113–7.

10. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med. 2018 Aug 7;18(3):91–3.

11. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63.

12. McDowell I. Measuring Health: A Guide to Rating Scales and Questionnaires. Oxford University Press; 2006. 765 p.

13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159–74.

14. Aldridge V, Dovey T, Wade A. Assessing Test-Retest Reliability of Psychological Measures: Persistent Methodological Problems. European Psychologist. 2017 Jan 1;22:207–18.

15. Massey Jr. FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association. 1951 Mar 1;46(253):68–78.

16. Wilcox RR. Some practical reasons for reconsidering the Kolmogorov-Smirnov test. Brit J Math & Statis. 1997 May;50(1):9–20.

17. Xie S, Wu J, Chen G. Discrete choice experiment with duration versus time trade-off: a comparison of test–retest reliability of health utility elicitation approaches in SF-6Dv2 valuation. Qual Life Res. 2022 Sep 1;31(9):2791–803.

18. Gamper E, Holzner B, King M, Norman R, Viney R, Nerich V, et al. Test-Retest Reliability of Discrete Choice Experiment for Valuations of QLU-C10D Health States. Value in Health. 2018 Mar 1;21.

19. Badia X, Monserrat S, Roset M, Herdman M. Feasibility, validity and test-retest reliability of scaling methods for health states: the visual analogue scale and the time trade-off. Qual Life Res. 1999 Jun;8(4):303–10.

20. Gudex C, Dolan P, Kind P, Williams A. Health State Valuations from the General Public Using the Visual Analogue Scale. Quality of Life Research. 1996;5(6):521–31.

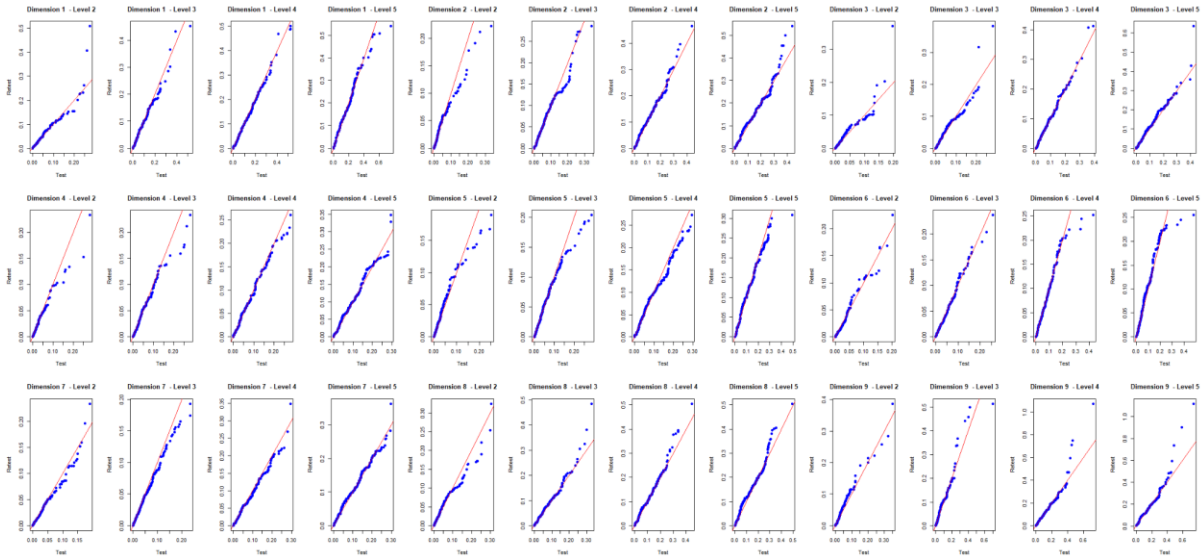
21. Jiang R, Shaw J, Mühlbacher A, Lee TA, Walton S, Kohlmann T, et al. Comparison of online and face-to-face valuation of the EQ-5D-5L using composite time trade-off. Qual Life Res. 2021;30(5):1433–44.

22. Norman R, King MT, Clarke D, Viney R, Cronin P, Street D. Does mode of administration matter? Comparison of online and face-to-face administration of a time trade-off task. Qual Life Res. 2010 May 1;19(4):499–508.

23. Robinson S. Test-retest reliability of health state valuation techniques: the time trade off and person trade off. Health Econ. 2011 Nov;20(11):1379–91.

APPENDIX

Figure S1. Q-Q plots based on the empirical distributions of the aggregate level utility decrements in the test and retest (Total sample)



Dimensions: 1, Mobility; 2, Daily activities; 3, Exhaustion; 4, Loneliness; 5, Cognition; 6, Anxiety; 7, Sadness/depression; 8, Control; 9, Pain